Michael Gilead
Kevin N. Ochsner *Editors*

The Neural Basis of Mentalizing

Springer

The Neural Basis of Mentalizing

Michael Gilead • Kevin N. Ochsner
Editors

The Neural Basis of Mentalizing

Springer

Editors

Michael Gilead
Psychology Department
Ben-Gurion University of the Negev
Beer Sheva, Israel

Kevin N. Ochsner
Department of Psychology
Columbia University
New York, NY, USA

ISBN 978-3-030-51889-9 ISBN 978-3-030-51890-5 (eBook)
https://doi.org/10.1007/978-3-030-51890-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book, *The Neural Basis of Mentalizing*, comes out 20 years after the publication of Chris Frith and Uta Frith's (1999) groundbreaking *Science* paper "Interacting Minds—A Biological Basis," which described the first major efforts to understand the neural bases of humans' ability to *mentalize*, namely, to reflect on the mental states of the self and of others.

In the 20 years that have passed, the study of mentalizing has continued to mature and has become a central topic in the world of cognitive psychology and cognitive neuroscience. Moreover, mentalizing research plays a crucial role in such diverse fields as social and developmental psychology, emotion research, clinical psychology, linguistics, game theory, artificial intelligence, philosophy of mind, and primatology. The current book provides a comprehensive collection of perspectives on the topic and brings together researchers from these diverse fields.

We hope and believe that the publication of this volume will give rise to further progress in the study of the vital human capacity to understand themselves and others.

We wish to thank all of the numerous contributors to this book, Springer-Nature, and the production team for all of their hard work.

Beer Sheva, Israel — Michael Gilead
New York City, NY, USA — Kevin N. Ochsner

Contents

Part I Introduction

Part II The Boundaries of Mentalizing

Part III Theoretical Approaches

Part IV Cognitive Components

Part V Mentalizing in Social Interactions and Decision-Making

Part VI Mentalizing in Self-Referential Processing and Emotion

Part I
Introduction

A Guide to the Neural Bases of Mentalizing

Michael Gilead and Kevin N. Ochsner

Mentalizing is the act of thinking about mental states, i.e., reflecting on one's own mental states and estimating the mental states of others. This capacity plays a crucial role in daily life, and it is widely believed that our advanced mentalizing abilities may be one of the main elements that distinguish us from other animals. As such, much research in psychology, philosophy, and neuroscience has been devoted to studying this process.

As implied by its title, the current book reviews the extant research examining mentalizing at the *neural* level; however, as suggested by Marr (1982), a full understanding of a cognitive process also entails an understanding the *algorithms* that are instantiated by the neural tissue (i.e., the representations and cognitive processes involved in the process), which in turn entails an understanding of the *computation* subserved by these processes (i.e., the challenge that the process is meant to address). In light of this, the 34 chapters in this book provide an analysis of mentalizing across the neural, algorithmic, and computational levels (Fig. 1).

This book comes out 20 years after the publication of Frith and Frith's (1999) influential paper describing the first major efforts to understand the neural bases of the mentalizing capacity. The main findings of these initial studies were that a network of temporo-parietal, anterior temporal, posterior medial parietal, and medial prefrontal regions subserve humans' mentalizing ability. Twenty years later, these findings have been replicated numerous times. Celebrating the 20 years' anniversary of this paper, the current volume begins with a special chapter by Frith and Frith, providing a historical perspective on the emergence of mentalizing research, an overview of current research, and an outline of future directions.

M. Gilead (✉)
Psychology Department, Ben-Gurion University of the Negev, Beer Sheva, Israel
e-mail: mgilead@bgu.ac.il

K. N. Ochsner (✉)
Department of Psychology, Columbia University, New York, NY, USA
e-mail: ko2132@columbia.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_1

Fig. 1 The most widely used terms throughout the 33 chapters of the book

Before reviewing the contents of the book, we wish to provide readers with a short guide to some of the theoretical terms and common anatomical labels that are widely used throughout.

Terminological Guide

There are some terms that simply overlap with the term Mentalizing. For example, we see no meaningful difference between the term Mentalizing and terms such as **Mindreading** and **Mental State Inference** or **Mental State Attribution**.

The psychological and neuroscientific literature often interchanges the term **Theory of Mind Reasoning** with the term mentalizing. As noted in the Chapter by Perner et al., a caveat of this is that it may be taken by some to suggest a commitment to a specific theory of how mentalizing takes place (namely, the *Theory Theory* of mentalizing, see chapters in Part II)—which one may or may not endorse.

As noted, to mentalize is to reflect about one's own mental states and estimate the mental states of others. If John slaps Jack, and as a result, Jack becomes angry—it is not necessarily the case that Jack engaged in mentalizing. Only when Jack reflects upon his feelings following the slap or tries to understand why John slapped him, then it is appropriate to say that mentalizing took place. Because mentalizing involves thinking about thinking (or similarly, thinking about emotions, desires, intentions, and so on), mentalizing is often referred to as **Meta-Representational** processing (as "meta-representation" means a representation of a representation).

In order to assess whether the behavior of an individual is indeed driven by Meta-Representation, the central paradigm used in Mentalizing research is the False-Belief paradigm (discussed at length throughout the book). This paradigm assesses whether individuals utilize meta-representations by examining behavior in a situation wherein the participant's representation of the world differs from that of another person. In light of this, the capacity examined on False-Belief tasks is often referred to as **False-Belief Reasoning**. False-Belief reasoning entails mentalizing and provides evidence for Meta-Representation; however, the terms are not interchangeable (False-Belief reasoning is mentalizing but not vice versa).

As reflected in the False-Belief paradigm, it is often the case that in order for mentalizing to be adaptive, individuals need to see the world from a perspective that differs from their own. The cognitive process of trying to view and interpret the world from the perspective of another person, rather than the self, is termed **Perspective Taking**. This process is believed to require the ability to overcome "egocentric biases" or "decouple the representation of the world as seen by myself and another." The terms Mentalizing and Perspective Taking are often used interchangeably. However, again, it is worth noting that individuals may engage in mentalizing without taking the perspective of another person (despite the fact that this would be maladaptive).

Further confusion is entailed by the fact that the term Perspective Taking also refers to a specific *strategy*, widely studied in social psychology (i.e., the explicit strategy of trying to understand another person by imaging oneself in their predicament). Additionally, the capacity to see the world through the perspective of another person does not necessarily involve mentalizing; for example, it is (at least theoretically) possible that tasks of simple visual Perspective Taking rely on mechanisms that do not involve the sort of thinking about thinking which we refer to as mentalizing.

As noted, the term Perspective Taking refers to the *process* whereby individuals try to see the world from a perspective that differs from one's own. It is assumed that such attempts are often successful and generates an accurate understanding of others' perspectives. In light of this Perspective Taking is often discussed as a *skill*. Similarly, the ability to accurately figure out the contents of another's mind is often termed **Empathic Accuracy** (see discussion in the Chapter by **Hinnekens**, **Ickes**, **Berlamont**, **and Verhofstadt**).

This brings us to another loaded term that partially overlaps with the term mentalizing (and is probably the most widely used in lay discourse), namely, **Empathy**. Just like Mentalizing, Empathy entails an understanding of the mental states of others; however, unlike mentalizing, when we say that John empathized with Jack, it typically means that John shared some aspects of the emotional experience of Jack. Researchers of empathy often speak of Emotional Empathy and **Cognitive Empathy**, with the former entailing experience sharing, and the latter concept being virtually identical to mentalizing.

Finally, people sometimes use the term **Social Cognition** interchangeably with the term mentalizing. The capacity for mentalizing is indeed part of the broader class of Social-Cognitive capacities, but there are various social-cognitive capacities that should not be thought of as mentalizing (see **Malle's** chapter). The term Social Cognition is especially confusing because it also refers to a field of research (namely, social psychological research that adopts the premises and toolkit of cognitive science).

Neuroanatomical Guide

Throughout the book, there are various different anatomical regions that are mentioned with regard to the mentalizing capacity; first and foremost of these is the Mentalizing Network. The Mentalizing Network is comprised of regions in the posterior temporal lobe and parietal lobe, medial prefrontal cortex, medial parietal cortex, and the anterior temporal lobe. This network overlaps with the so-called Default Mode Network (Raichle et al., 2001). However, under some definitions of the Default Network, it includes several areas that extend beyond the classic Mentalizing Network. Below, we provide a quick neuroanatomical guide to the approximate locations of the regions discussed in the chapters.

Throughout the book, crucial parietal and posterior temporal areas for mentalizing are referred to as the pSTS (posterior superior temporal sulcus), TPJ (temporoparietal junction), Angular Gyrus, and IPL (inferior parietal lobule). In Fig. 2, we used the automated meta-analysis tool Neurosynth (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011) to show the peak coordinates for each of these anatomical labels, as well as the peak posterior temporal and parietal locations of the mentalizing network and default-mode-network.

Crucial medial prefrontal areas for mentalizing are often referred to in the book as the vmPFC (ventromedial prefrontal cortex), dmPFC (dorsomedial prefrontal cortex), and simply mPFC (medial prefrontal cortex). In Fig. 3, we show the peak coordinates for each of these anatomical titles, as well as the peak medial prefrontal coordinates of the mentalizing network and default-mode-network.

Crucial medial parietal areas for mentalizing are referred to in the book as PCC (posterior cingulate cortex) and Precuneus. In Fig. 4, we show the peak coordinates for each of these anatomical titles, as well as the peak medial parietal coordinates of the mentalizing network and default-mode-network.

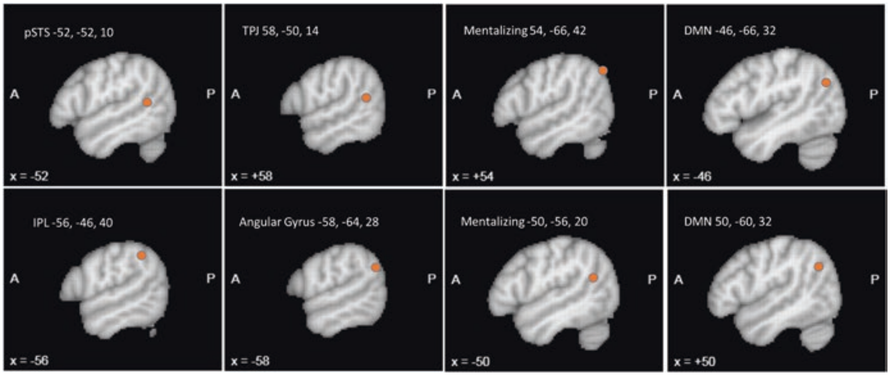


Fig. 2 Peak Neurosynth coordinates of main parietal and posterior temporal regions discussed throughout the book. Also shown are peak temporal and parietal coordinates for the terms Mentalizing and Default Mode Network

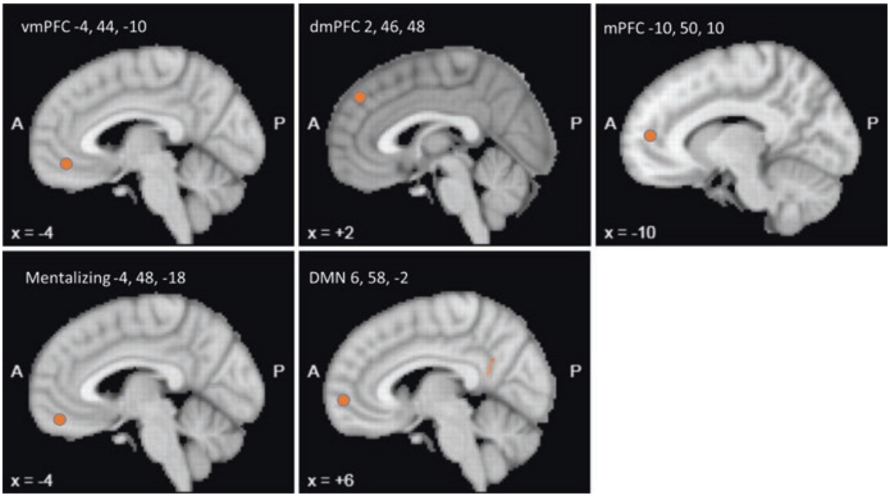


Fig. 3 Peak coordinates of medial prefrontal regions discussed throughout the book, and medial prefrontal peaks of the Mentalizing Network and Default Mode Network

Finally, a crucial anterior temporal region in the mentalizing network is often referred to in the book as the Temporal Pole. In Fig. 5, we show the peak coordinates for this anatomical title, as well as the peak anterior temporal coordinates of the mentalizing network.

Chapter Guide

As noted, the current volume begins with a special 20 years' anniversary chapter by **Frith and Frith**, providing a historical perspective on the emergence of mentalizing research, an overview of current research, and an outline of future directions. The subsequent chapters are organized into five parts: (i) The boundaries of mentalizing; (ii) Theoretical approaches to the study of mentalizing; (iii) The components of mentalizing; (iv) Mentalizing in social interactions and decision-making; (v) Mentalizing in self-referential processing and emotion.

Part I: The Boundaries of Mentalizing

One of the ways by which scientists can study a phenomenon is by carefully mapping its boundaries. In the case of mentalizing such research has attempted to examine the extent to which humans are able to accurately infer the mental states of others and outline the degree to which different populations (e.g., infants,

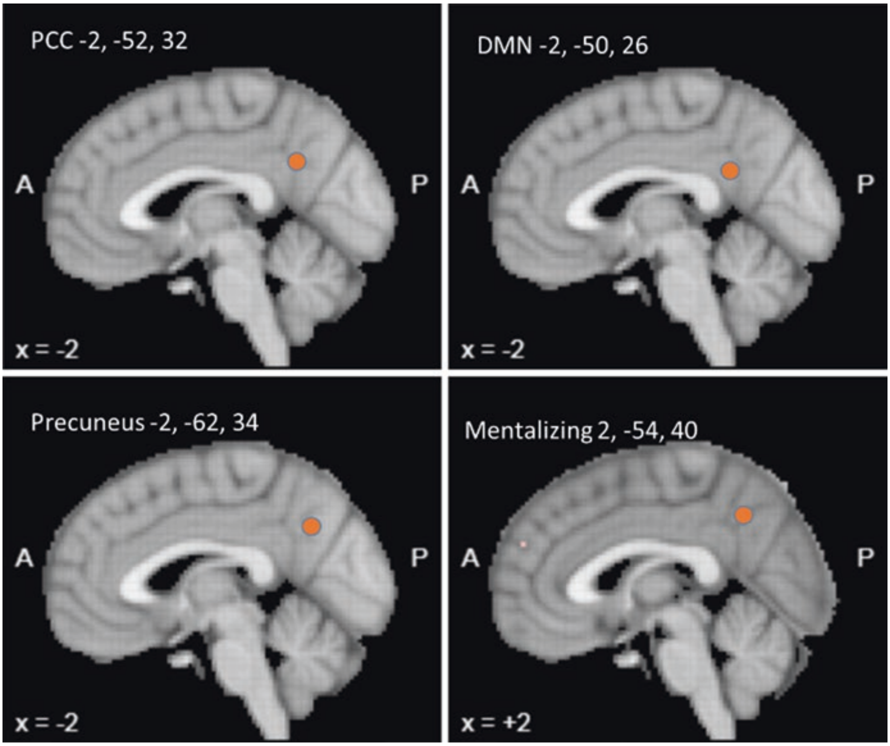


Fig. 4 Peak coordinates of posterior medial parietal regions discussed throughout the book, and posterior medial parietal peaks of the Mentalizing network and Default Mode Network

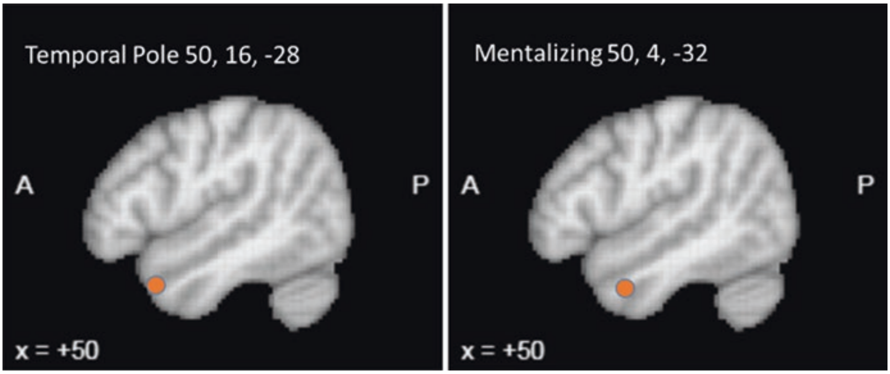


Fig. 5 Peak coordinates of anterior temporal regions discussed throughout the book, and anterior temporal peaks of the Mentalizing network

neuroatypical individuals) and different species (e.g., nonhuman primates) exhibit mindreading capacities. Thus, the first part of the book describes research that maps the mentalizing ability by establishing its boundaries.

As reviewed in the chapter by Frith and Frith, the extensive research program on humans' capacity for mental state processing can be largely traced back to a highly influential paper, by Premack and Woodruff (1978), that asked the question of whether nonhuman primates have an explicit "theory of mind." As a response to this paper, several philosophers suggested a litmus test to determine whether an organism indeed represents the mental states of others—a test subsequently termed the false-belief task. Several years later Wimmer and Perner (1983) devised a false-belief task and ran it on young children and discovered that it is only by the age of 4 that children are able to perform this task successfully.

Following the finding that children fail to pass false-belief tasks before the age of 4, an additional crucial milestone in mentalizing research came from the discovery that children with Autism Spectrum Disorder likewise often fail false-belief tasks (Baron-Cohen et al., 1985). Such findings have led to an extensive research program on impairments in mentalizing capacities in neurotypical and neuroatypical populations. In their chapter, **Simantov**, **Lombardo**, **Baron-Cohen**, **and Uzefovsky** discuss one specific process that may be impaired in neuroatypical individuals and may play a crucial role in subserving the capacity to mentalize—namely, the capacity for *self-other distinction*. An alternative view of mentalizing impairments in Autism is discussed by **Herrington**, **Parish-Morris**, and **Schultz**, implicating decreased social motivation in neuroatypical individuals.

Whereas the research program initiated by Wimmer and Perner's (1983) false-belief paradigm research suggested that a full-blown theory of mind does not develop until the age of 4, further research has suggested a dissociation between infants' explicit responses on false-belief tasks and their behavior. This research (e.g., Onishi & Baillargeon, 2005; Southgate, Senju, & Csibra, 2007) showed that eye-movement patterns of infants as young as 2 years old might reflect advanced implicit mentalizing capacities. In recent years, these results and their interpretation have been hotly debated; in their chapter, **Grosse Wiesmann and Southgate** provide an attempt towards an integration of supposedly incongruent past findings—that arrives at the interesting conclusion that while infants may not have fully fledged mentalizing capacities, they may represent the world in an altercentric manner; namely, the default condition for humans may be to see the world as others see it.

In a nice example of a long scientific journey, the eye-tracking measures developed for the study of mentalizing ability in young children provided a breakthrough in research into Premack and Woodruff (1978) question concerning the existence of theory-of-mind in nonhuman primates. After almost 40 years of research, Krupenye, Kano, Hirata, Call, and Tomasello (2016) utilized eye-tracking to provide compelling evidence that, indeed, chimpanzees seem to pass the false-belief tasks, and as such may have meta-representational capacities. In the current volume, **Krupenye** discusses this research and what it means for questions regarding the evolution of Mentalizing. In contrast, **Arre and Santos** voice a more pessimistic view of the mentalizing capacities of nonhuman primates and suggest that further research is warranted in order to rule out the possibility that mentalizing is a uniquely human capacity.

Valuable insights concerning mentalizing and its limitations come from behavioral research on neuro*typical* adults. For example, much research has utilized paradigms gauging participants' empathic accuracy (Ickes, 1993) to quantify the extent to which individuals are able to correctly read the minds of others. The chapter by **Hinnekens**, **Ickes**, **Berlamont**, **and Verhofstadt** reviews the rich literature and the consequence of empathic accuracy to social interaction and well-being. **Jhurry and Harris** review the contextual factors that determine when and why individuals fail to mentalize and the different implications of reduced mentalizing. Finally, the chapter by **Atias and Aviezer** evaluates the extent to which people are able to infer mental states from facial expressions and vocalizations. They highlight how, contrary to popular belief, it is often nearly impossible to gauge people's emotions based purely on such nonverbal information.

Part II: Theoretical Approaches to the Study of Mentalizing

Throughout the years several prominent theoretical perspectives have attempted to explain the mechanisms by which mentalizing takes place. The chapters in Part II review these different scientific approaches.

Houlihan, **Tenenbaum**, **and Saxe** present an overview of their extant work on computational modeling of the mentalizing capacity with the "Bayesian Theory of Mind" approach, as well as their research on the type of information represented within the mentalizing network. They discuss their view according to which the computations performed by the mentalizing network can be best described as operations that occur over structured representations of mental content, wherein the different elements stand in specified relations to each other—just like in a scientific *theory* (Carey, 2009; Gopnik & Wellman, 1994).

As Houlihan et al. note, this approach, which describes mental state inference as in terms of structured relations between representations of mental states and behaviors, stands in contrast to the recent computational model proposed by Tamir and Thornton. This approach, described in the chapter by **Thornton and Tamir**, suggests that we infer people's mental states by forming a low-dimensional representation of mental states, traits, and actions—and the transition probabilities between different points along this representation. The findings described in Thornton and Tamir's chapter provide evidence that the mentalizing network may indeed represent the mental world by relying on nonstructured associations between elements in a low-dimensional space, thus challenging the structured representation approach. In contrast, findings reported in Houlihan et al. challenge the idea that a low-dimensional representation suffices to explain the variance in mental state attributions.

Such conflicting models may provide a roadmap for empirical research into the computational and neural bases of mentalizing. Furthermore, as reviewed in **Gonzalez and Chang's** guide to computational models of mentalizing, in recent years, several additional computational models provide researchers with viable alternatives, and that will likely guide future empirical investigation in the field.

The approach described by Houlihan et al. wherein mentalizing is achieved via statistical inference processes upon structured meta-representations is a reincarnation of the classic philosophical view termed the theory-theory of mental state inference. An article published in 1986 by Robert Gordon provided compelling challenges to the theory-theory approach and instead argued for what is termed the *Simulation* theory. Simulation theory suggests that in order to understand an organism like myself, I do not have to generate a theory of myself—but rather simply examine my own responses in similar situations. In his chapter, **Gordon** discusses the simulation view as a natural consequence of recent computational models of cognitive processing, namely, the *predictive processing* approach. Furthermore, Gordon argues that the computational toolkit utilized by recent incarnations of the theory-theory (i.e., the *probabilistic generative modeling* approach) naturally invites the view that mental state inference does not involve operations upon a model of mental states, but rather an inversion of one's own mechanisms for action selection.

One of the classic arguments against the simulation approach is that the self is oftentimes an inappropriate model for thinking about the mental states of others. Like Gordon, **Perner**, **Aichhorn**, **Tholen**, **and Schurz** suggest that people nonetheless start off with an assumption of identity between self and other. The *Teleological* theory, described by Perner et al., presents a stark alternative to specialized accounts of mental state inference and argues that mentalizing happens via reliance on general-purpose cognitive processes that we use in order to think about nonmental content. Specifically, the ability to think about reasons for action and the ability to find the identity between different senses of the same referent object.

While the different perspectives presented in this chapter significantly differ from each other, they are all comfortably situated within the realms of the *information processing approach*; namely, the idea that "the mind is a system of organs of computation, designed by natural selection to solve the kinds of problems our ancestors faced in their foraging way of life" (Pinker, 1997, p. 21, cited in Carpendale et al.). **Carpendale**, **Müller**, **Lewis**, **and Wallbridge** suggest that explaining mentalizing (and cognition more broadly) by using an ontology of such "organs of computation" is the wrong idea. Instead, they present a perspective on mentalizing they term the "process-relational" account, which focuses on understanding how children construct meaning through an interaction with their physical and social world.

Part III: The Components of Mentalizing

In contrast to the dissenting view championed by Carpendale et al., the way cognitive scientists have typically attempted to explain the process of mentalizing is by breaking this process down to its constituent computational components, and trying to understand how these components interact in the mind and brain. In Part III we discuss the research into these purported cognitive components and how they are implemented in the brain.

In order to provide a basis for such a componential view, **Malle** contextualizes the process of mentalizing within the broader realm of social-cognitive capacities and proposes a hierarchically organized set of different computations that allow humans to understand the minds of others. This model of a "tree of social cognition" provides an ontology of the components of social cognition and their relation to each other. **Apperly** describes a process model that explicates the temporal progression of the cognitive components involved in mentalizing. Specifically, Apperly proposes a dual-system approach to mindreading that constitutes a slow-but-flexible processing stream that heavily relies on retrieval from long-term memory and on cognitive control and a fast-but-inflexible stream that entails lower cognitive demands.

The chapter by **Van Overwalle and Heleven** provides an in-depth review of the neuroscientific research into many of the component processes discussed in Apperly's process model. Specifically, the chapter reviews extant research on the neural basis of the "slow" and "fast" social cognition and on the cognitive control of memory during mentalizing. Furthermore, they describe a theory of how different regions of the Mentalizing Network play a role in these different component processes.

While *social*-cognitive neuroscientists often study the Mentalizing Network in the context of social cognition tasks, this network has also been the focus of much research by *cognitive* neuroscientists that have attempted to characterize the broader (and potentially non-social) underlying computations of this network of regions, also termed the Default Mode Network.

For example, research into semantic cognition has revealed that the neural bases of semantic processing overlap with the Default Network (Binder, Desai, Graves, & Conant, 2009). In light of this, in his chapter, **Binder** proposes that researchers should think of mentalizing-related activity as a manifestation of conceptual processing; namely, that mentalizing is simply an application of humans' more general capacity for conceptual thought in the social domain. Relatedly, in her chapter, **de Villiers** surveys the research that examined the role of language processing in mentalizing and considers the evidence for several views regarding the relation mentalizing conceptual and linguistic processing.

Whereas much research suggests the Default Network is associated with *semantic* processing and semantic memory, this network is also widely implicated in supporting *episodic* memory (e.g., Schacter, Norman, & Koutstaal, 1998). In their chapter, **Van Genugten and Schacter** review the evidence that parts of the DMN subserve the retrieval of specific event details and explain how such retrieval processes can be used to simulate the content of another person's mind and their future behavior.

Finally, given the involvement of the DMN in both semantic and episodic memory, **Baror**, **Aminoff**, **and Bar** present their view of this network as subserving a potentially more fundamental process of context-based associative processing.

They explain how such contextual associations may play a role in semantic cognition, episodic memory, and mentalizing.

Part IV: Mentalizing in Social Interaction and Decision-Making

A comprehensive understanding of the process of mentalizing entails understanding its component processes, neural basis, but also, importantly, the role of this process in adaptive human behavior. In recent years, much research by social-cognitive neuroscientists has elucidated how mentalizing subserves social interaction and decision-making. In Part IV, the contributors review these research endeavors.

It is widely argued that the evolutionary success of *Homo sapiens* relied on our ability to cooperate with each other and accumulate cultural knowledge (Boyd, Richerson, & Henrich, 2011). This unique human capacity is predicated on our ability to effectively communicate with each other and on our motivation to help each other out (Tomasello, Carpenter, Call, Behne, & Moll, 2005). As discussed in the chapter by **Parelman**, **Doré**, **and Falk**, neuroscientific research highlights the importance of mentalizing processes and of the mentalizing network in communication behaviors. Additionally, **Franklin-Gillette and Shamay-Tsoory** discuss the neuroscientific literature on empathy and its role in prosocial behavior such as providing emotional support for others. These two chapters explain how mentalizing lies at the heart of our species' highly interdependent, cooperative and helpful nature.

To survive in the evolutionary arms race, organisms developed mechanisms that help them attain survival-related goals. Within the world of decision science, the attainment of survival-related goals is described as providing organisms with some *utility* or *value*; decision scientists try to understand the mechanisms that subserve utility or value-seeking behavior. As highlighted by **Charpentier and O'Doherty**, because humans are fundamentally social creatures, attainment of value often occurs with the help of others (e.g., observing their behaviors to understand which actions yield positive consequences) or by competing with others (e.g., in economic negotiations)—and as such, relies on mental state inferences. In their chapter, they describe neuroscientific research that utilized computational modeling of decision-making, explicating the role of mentalizing in such decision process. Whereas Charpentier and O'doherty primarily focus on reward seeking, **Espinosa, Golkar and Olsson** mainly discuss research on learning from the behaviors of others with regard to punishments and highlight how such learning plays a crucial role in shaping our affect experience and decisions—and the role of mentalizing in such processes.

As highlighted by **Civai and Sanfey**, because our lives are embedded within social contexts, many decisions we make do not only affect material rewards and punishments but also influence our standing in the social world. In their chapter, they review neuroeconomic research showing how social considerations such as prosocial motivation, fairness, and reputation affect decision-making processes;

moreover, they discuss how our decision-making processes are shaped by our expectations of others' social behavior.

A fundamental process in forming such expectation involves generating an impression of people's stable dispositions; as argued by **Ray**, **Mende-Siedlecki**, **Gantman**, **and Van-Bavel**, the main factor in dispositional inferences is an evaluation of a person's moral character. The findings discussed in these two chapters highlight the involvement of mentalizing network in shaping these important social decisions and inferences.

One of the assumptions of decision science is that human beings are rational creatures (i.e., the so-called "homo-economicus" perspective). A significant challenge to this economic perspective to human behavior is human cognition is often biased and appears to be suboptimal. For example, in social decision-making people often perceive reality as they want it to be (e.g., everybody loves me), rather than how it truly is. In the final chapter of this part, **Park**, **Kim**, **and Young** discuss the notion of rational choice and suggest that activity in the mentalizing network may be a marker of procedurally rational behavioral (i.e., decision that is consistent with Bayesian Decision Theory). Together, the findings reviewed in Part IV highlight how understanding the process of mentalizing is crucial for understanding other vital aspects of human behavior.

Part V: Mentalizing in Self-Referential Processing and Emotion

In the final part of the book, the authors discuss how mentalizing shapes our sense of self and our emotional experience—and thereby how it plays a role in psychopathology.

According to the symbolic interactionist perspective (Mead, 1934), our mental representation of the self is construed from representations of other people and of the way they see us—and as such, the sense of self depends on the process of mentalizing. Indeed, as reviewed in the chapter by **Chavez**, the neuroscientific research on representation of self and others shows that the neural basis of self- and other-related processing is highly interdependent.

It has been suggested (e.g., Leary & Baumeister, 2000) that the very reason that we need a sense of self is because it allows us to keep track of how others view us (e.g., whether they think we are valued members of society), and thereby assess our social standing, and the risks of being ostracized by our peers. Because of this, perceptions of the self are intimately related to our affective experience and well-being.

Much research, reviewed by **Sahi and Eisenberger**, suggests that the feeling of being negatively evaluated by others—which relies on process of mentalizing—is literally a painful experience. Moreover, in light of the potentially aversive nature of self-evaluation, it should be unsurprising that when self-evaluative processes become dysregulated, psychopathology ensues. Based on such reasoning, **Maresh and Andrews-Hanna** discuss how aberrations in self- and other-related processing can give rise to social anxiety disorder.

The idea that mentalizing processes play a central role in psychopathology has become central in the world of psychotherapy research and practice, wherein it is referred to as the *mentalization* approach (e.g., Fonagy & Bateman, 2008). This approach has given rise to one of the leading psychotherapeutic methods in modern clinical psychological science, namely, *mentalization-based treatment*. The premises of the rich theory and their relation to the neuroscientific findings on mentalizing are reviewed in a chapter by **Luyten**, **De Meulemeester**, **and Fonagy**.

Summary

More than 40 years after Premack and Woodruff (1978) and 20 years after Frith and Frith (1999), the study of mentalizing continues to blossom, with new theories and discoveries published every week. While there are many unanswered questions and controversies, it seems to us that the steady advances in research into this topic highlight that the study of mentalizing is one of the most successful enterprises in psychological science. The 33 chapters in the current volume provide what is likely the most comprehensive collection of perspectives published to date on the topic. Hopefully, the publication of this volume will prompt further advances in the study of the fundamental and crucial human capacity to understand themselves and others.

References

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex, 19*(12), 2767–2796.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind”? *Cognition, 21*(1), 37–46.

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences, 108*(Suppl 2), 10918–10925.

Carey, S. (2009). *The Origin of Concepts*. New York: Oxford University Press.

Fonagy, P., & Bateman, A. (2008). The development of borderline personality disorder - A mentalizing model. *Journal of Personality Disorders, 22*(1), 4–21. https://doi.org/10.1521/pedi.2008.22.1.4

Frith, C. D., & Frith, U. (1999). Cognitive psychology - Interacting minds - A biological basis. *Science, 286*(5445), 1692–1695. https://doi.org/10.1126/science.286.5445.1692

Gopnik, A., & Wellman, H. M. (1994). The theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (p. 257). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511752902.011

Ickes, W. (1993). Empathic accuracy. *Journal of Personality, 61*(4), 587–610. https://doi.org/10.1111/j.1467-6494.1993.tb00783.x

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science, 354*(6308), 110–114. https://doi.org/10.1126/science.aaf8110

Leary, M. R., & Baumeister, R. F. (2000). The nature and function of self-esteem: Sociometer theory. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 32, pp. 1–62). San Diego, CA: Elsevier Academic Press Inc..

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information* (2nd ed.). New York, NY: Henry Holt and Co. Inc.

Mead, G. H. (1934). *Works of George Herbert Mead: Vol. 1. Mind, self, and society: From the standpoint of a social behaviorist*. Chicago, IL: University of Chicago Press.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255–258. https://doi.org/10.1126/science.1107621

Pinker, S. (1997). *How the mind works*. New York, NY: W. W. Norton & Company

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(04), 515–526.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences, 98*(2), 676–682.

Schacter, D. L., Norman, K. A., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology, 49*, 289–318. https://doi.org/10.1146/annurev.psych.49.1.289

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18*(7), 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 675. https://doi.org/10.1017/s0140525x05000129

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs - Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods, 8*(8), 665–U695. https://doi.org/10.1038/nmeth.1635

Mapping Mentalising in the Brain

Chris D. Frith and Uta Frith

Introduction

We were delighted to have been asked to revisit our paper *Interacting minds—a biological basis* (C. D. Frith & U. Frith, 1999). Taking a personal perspective, looking back as well as speculating about the future, is an indulgence and a diversion when you get to our age. The great thing about being our age is that we feel free from pressures, such as the need to draw attention to our work or to demonstrate its value. We can therefore look at the various steps and missteps that we and others took in mapping out the field with a degree of equanimity that we might not have been able to muster before. We are happy to survey ideas and evidence contributed from different points of view, not exhaustively, but with the eyes of incurable enthusiasts.

A Novel Cognitive Ability

The term 'Theory of Mind' comes from the title of a paper by Premack and Woodruff (1978), 'Does the chimpanzee have a theory of mind?' This paper marks an ignition point for ideas on social cognition, whether from philosophy (D. C. Dennett, 1987;

C. D. Frith (✉)
Wellcome Centre for Human NeuroImaging, University College London, London, UK

Institute of Philosophy, School of Advanced Study, University of London, London, UK

UCL Institute of Cognitive Neuroscience, London, UK
e-mail: c.frith@ucl.ac.uk

U. Frith
UCL Institute of Cognitive Neuroscience, London, UK
e-mail: u.frith@ucl.ac.uk

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_2

Searle, 1995), animal behaviour (Byrne & Whiten, 1989; Woodruff & Premack, 1979), evolutionary psychology (Humphrey, 1976), linguistics (Sperber & Wilson, 1986) or cognitive development (Leslie, 1987; Wimmer & Perner, 1983). This was the first time that all these ideas all came together. The phrase 'having a theory of mind' started to be used so much that its abbreviation to ToM became necessary. It was in many ways an awkward label, but also a convenient way to refer to the assumption that we can reason about peoples' behaviour on the basis of their hidden mental states.

It was not self-evident what was meant by mental states. Mental states were specified by the examples of desires, intentions, beliefs, and knowledge, and added somewhat later, pretence and irony. All these mental states have in common that they are not free floating, like moods, but instead they are always 'about something' (Brentano, 1995/1874), and they have consequences for actions. For instance, a teacher's belief that a student has had a significant amount of help to complete his project is independent of whether or not this was actually the case. The teacher's belief will determine the mark the project will get. Explaining behaviour with reference to these types of mental states is the lifeblood of folk psychology. 'Why did Paul take a taxi when he is hard up?' 'Because he *believed* the busses were on strike, and he *intended* to keep his appointment'. This contrasts with the statement, 'he took a taxi because there was a strike on'. In this case, we are explaining his behaviour in terms of the physical state of the world.

There was clearly a whole package of ideas compressed into what is meant by 'having a theory of mind'. As these ideas spread, other phrases came into use as well, such as 'taking an intentional stance' (D. C. Dennett, 1987). Because there was a need for a verb to denote the action implied in this new cognitive ability, the term 'mentalising' was introduced (U. Frith, 1989). All these terms refer to the ability to make inferences about mental states.

The Importance of False Beliefs

In response to Premack and Woodruff's somewhat tongue-in-cheek question, a number of commentators (Bennett, 1978; Dennett, 1978; Harman, 1978) suggested that the acid test of ToM would be the ability to recognise that someone has a false belief and know how they will behave on the basis of that belief. Having a true belief was not enough of a test, because here the mental and physical states of affairs coincide. In the case of Paul taking a cab because he believed there was a strike—and at the same time there really was a strike—you cannot easily tell whether his behaviour is determined by his belief or by the actual state of the world at the time. This suggestion inspired Wimmer and Perner (1983) to develop a number of false belief tests, which were so beautifully simple that they could use them to find out at what age children showed evidence of a Theory of Mind. The first test they came up with is the classic Maxi task, where a story is enacted with the aid of simple props: Maxi puts his chocolate in the red cupboard and goes out to play. While he is out, his mother enters and moves his chocolate to the blue cupboard and leaves. Then Maxi comes back and the child is asked, 'where will Maxi look for his chocolate?' If they can mentalise, the children will reason that Maxi does not know that the chocolate has been moved and will therefore look in the red cupboard where he believes it to be. Wimmer and Perner found that children do not pass this false belief test reliably until 4–5 years of age. This has proved to be a remarkably robust finding (e.g. Wellman, Cross, & Watson, 2001).

What Autism Reveals About Mentalising

Shortly afterwards, the same test, in the version known as the 'Sally-Ann task', was given to autistic children (S. Baron-Cohen, Leslie, & Frith, 1985). This was not random coincidence. We were motivated by our then rather startling idea that the core social impairments of autism might be explained by a lack of ToM. Here was a way to test this hypothesis. If true, then autistic children should not be able to pass a simple false belief test. This was a novel prediction, and, amazingly, it was confirmed at first try. In line with Wimmer and Perner's findings, most of the control children, who had a mental age of around 5 years, passed the test. In line with our prediction, most of the autistic children failed, even though their chronological and mental age was higher. A striking feature of the result is that this difficulty with mentalising seemed to be rather circumscribed. It is not simply a general consequence of low IQ, since children with Down's syndrome could pass the test with a somewhat lower IQ than those with autism.

However, many more experiments with different paradigms were necessary to make a robust case. In particular, it was necessary to show that autistic children were able to pass tests of similar difficulty that did not involve reasoning about mental states. In a subsequent study, using a completely different task, the critical difference between the groups was found again (Simon Baron-Cohen, Leslie, & Frith, 1986). Here children had to order pictures so as to make up a story and talk about it afterwards. In one condition the story made sense only when mental states were inferred. Here the autistic children did much worse compared to both neurotypical children and children with Down's syndrome. However, they were at least as good as, if not better than the other children, in the contrasting condition, where physical states had to be inferred to make sense of the story.

Mentalising as a Cognitive Ability

We cannot emphasise enough that mentalising is a cognitive capacity underlying a range of behaviours, and not simply performance on a test. Whether a child passes or fails the Maxi/Sally-Ann test could be due to a number of reasons. However, the test was easy to demonstrate and served as a means of popularising a novel concept.

Hence it came to serve as an operational definition of mentalising. On the one hand, this helped it to gain credibility; on the other hand, it gave the misleading impression that mentalising was a behaviour, when actually it is a hidden cognitive ability. Such abilities can be tapped by behaviour but they are not the same as the behaviour. Task performance is always influenced by a variety of factors, including the participant's temporary state of motivation and attentiveness. Only by stripping away the specific mentalising component of the task from other components can we gauge what kind of entity it is. This demands careful design by making fine cuts in task comparisons (U. Frith & Happé, 1994) and a comparison of behaviour across many different tasks. This is a work in progress.

A Neurocognitive Basis

The findings using the fine cuts approach suggested to us that mentalising might be associated with a rather specific cognitive mechanism, which we believed must be underpinned by a circumscribed brain system. Predictably, most of our colleagues thought this was a very foolish idea. Surely a concept, such as mentalising, was far too abstract and complex to be associated with a circumscribed brain system. We took no notice. What partly drove us was the hope that this brain system would reveal structural or functional anomalies in autistic brains. But this turned out to be a hope too far. In fact, this search is still ongoing, but the problem is that we do not know which is the appropriate level to look, e.g. synapse, cells, connections, and circuits?

We were fortunate at that time to have access to early brain imaging equipment (positron emission tomography, PET) and conducted a study in which a few brave adult volunteers were scanned while reading stories which either required mental state reasoning (about false beliefs) or physical state reasoning (P. C. Fletcher et al., 1995). The volunteers had to be brave, because of the unavoidable risk of radiation in PET scanning. In the same year, Goel, Grafman, Sadato, and Hallett (1995) reported a similar study in which people were scanned while performing another kind of mental state reasoning: they were shown various man-made artefacts and asked if Christopher Columbus would have been able to work out what they were for. Both studies revealed a circumscribed set of brain regions that were activated when people had to think about the mental states of others, in particular, medial frontal cortex (mPFC), superior posterior temporal sulcus (pSTS) and posterior cingulate cortex (PCC).

By today's standards these studies were grossly underpowered ($n = 6$ and $n = 9$ respectively), but remarkably this network of regions has been repeatedly identified in studies of mentalising (see, e.g. Van Overwalle, 2009) and is sometimes loosely referred to as the social brain. The possibility of scanning many more volunteers and using many more and better controlled studies was greatly facilitated by the changeover from PET scanning to functional Magnetic Resonance Imaging (fMRI).

In discussing the findings, we have deliberately chosen to be somewhat vague about the precise locations of the three main components of the mentalising system. This is because our ability to localise brain regions far outstrips our models of brain function in relation to mentalising. Thus, *mPFC* refers to a large area of medial prefrontal cortex (Brodmann areas medial 9 and 10) and may include area BA 32 which is in anterior cingulate cortex rather than prefrontal cortex. We adopt similar prevarication by referring to *TPJ/pSTS* and *precuneus/PCC*. The precuneus is medial parietal cortex (BA7m), while PCC is retrosplenial cortex (BA 30 and 31). Perhaps, the most interesting goal for future research is to discover precise details about the connections between these landmarks (Mars, Sallet, Neubert, & Rushworth, 2013). We like to think of the mentalising system as a well-engineered circuit, because the components need to work together. It is encouraging that the connections are beginning to be revealed (see, e.g. Wittmann, Lockwood, & Rushworth, 2018).

The Social Brain

The notion of a ‘social brain’ was mooted earlier by Leslie Brothers (L Brothers, 1990). Brothers observed that neurons in the amygdala of awake behaving monkeys would respond to specific complex social stimuli, such as the approach of another monkey (Leslie Brothers, Ring, & Kling, 1990). She proposed that there was a brain system specialised in perception of social stimuli, such as faces and their expressions, or gestures and movements, from which the intentions and motivations of others could be inferred. This led to new research into the perception of such stimuli.

Biological motion was another important social stimulus that was first investigated using brain scanning at that time. Human motion kinematics, even if reduced to a handful of dots, are sufficient for observers to recognise the actions being performed. Observation of such motion elicits activity in pSTS (Bonda, Petrides, Ostry, & Evans, 1996; Jellema & Perrett, 2003; Puce, Allison, Bentin, Gore, & McCarthy, 1998). Even more specific activity, related to the observation of action, is that seen in mirror neurons, a striking phenomenon also first observed in monkeys at this time in Rizzolatti’s lab (Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992). These neurons fire when a monkey performs a particular action (e.g. a precision grip) and also when the monkey sees someone else performing the same action. The existence of such neurons requires that the brain has solved the correspondence problem relating what we see to what we do (Brass & Heyes, 2005). The location of mirror neurons in the motor system rather than the perceptual system provides an important clue as to how the correspondence problem is solved. We make inferences about the goals of the movements of others using our own motor system (J. M. Kilner, Friston, & Frith, 2007). The human mirror neuron system, sometimes called the action observation network (AON), consists of ventral premotor cortex, inferior parietal lobule and superior temporal sulcus (James M. Kilner, 2009).

Homing in on the Mentalising System

At the time we wrote our paper on interacting minds (C. D. Frith & U. Frith, 1999), it seemed plausible that the regions highlighted in these various studies were all part of a complex brain system specialised for processing social information. This included several cognitive capacities relevant to interactions with others: it could detect biological motion, recognise emotional expressions, imitate the actions of others, and make inferences about the mental states of others. This seemed to be validated by autism, because individuals with autism were thought to be impaired in all of these abilities. Indeed, there was the proposal that the mirror system might be the origin of all social impairments in autism (e.g. Ramachandran & Oberman, 2006). However, it turned out that autistic people can and do imitate actions (Fan, Decety, Yang, Liu, & Cheng, 2010; Hamilton, Brindley, & Frith, 2007). Furthermore, Marsh and Hamilton (2011) found evidence for a dissociation: autistic individuals showed atypical neural responses in the medial prefrontal mentalising system, but not in the parietal mirroring system.

What is the relationship between mentalising and mirror systems in the social brain? It seems they have complementary roles (de Lange, Spronk, Willems, Toni, & Bekkering, 2008; Grèzes, Frith, & Passingham, 2004; Redcay & Schilbach, 2019) and, in some contexts, there is evidence of an antagonistic relationship. For example, action observation leads to the automatic activation of the mirror neuron system, while the mentalising system is deployed when this automatic imitation has to be suppressed (Spengler, von Cramon, & Brass, 2009). Antagonism between the mirror system and the mentalising system is also suggested by the observation that people are better able to detect deception if they supress their tendency to imitate the gestures of the deceiver (Stel, van Dijk, & Olivier, 2009).

We conclude from the research in both neurotypical and autistic populations that it is important to distinguish mentalising from other social processes such as mirroring. Therefore, in what follows we shall adopt a framework in which the mentalising system is treated as a distinct, specialised system within the social brain.

Inputs to the Mentalising System

The mirror system responds rather specifically to the observation of goal-directed movements, although they do not need to be made by a living agent (Cross et al., 2012). In contrast, a surprising feature of the mentalising system is the range and abstract nature of the kinds of stimuli that elicit increased activity. For mentalising to be triggered, it is not necessary to see a real social interaction taking place. Reading a narrative about a social interaction, as used in our first PET study (P. C. Fletcher et al., 1995), as well as in many subsequent fMRI studies (e.g. R. Saxe & Kanwisher, 2003), is sufficient to activate the system. Subsequent studies have also shown that a wide variety of stimuli will have the same effect (see Schurz, Radua, Aichhorn, Richlan, & Perner, 2014 for a very useful review and meta-analysis). In an early PET study, participants viewed cartoon strips illustrating social interactions (Brunet, Sarfati, Hardy-Bayle, & Decety, 2000). In one of our own PET studies, participants watched animations, based on the classic demonstration of Heider and Simmel (1944). Pairs of triangles, viewed from above, moved about in ways which robustly elicited the perception of intentional behaviour. Such movements, which do not even resemble biological motion, activated the mentalising system (Castelli, Frith, Happe, & Frith, 2002).

A striking feature of all these paradigms is that they do not involve face to face social interactions. Rather, they require participants to observe other people interacting or answer hypothetical questions about what people might do in certain social situations. Schilbach et al. (2013) argued that to get a proper understanding of mentalising we should study what happens when people interact with one another in real time. This has been highlighted again recently by Wheatley, Boncz, Toni, and Stolk (2019).

There are indeed surprisingly few studies of this kind, typically using economic games. In one early study, people were scanned while playing the ultimatum game or prisoner's dilemma (Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004). At the moment when participants saw the offer that was made by their partner, and when they were presumably making inferences about their partner's intentions, activity increased in the mentalising system, in particular mPFC and pSTS. An interesting aspect of this study is that the behaviour of the opponents in the games was generated by a computer. However, in one condition, participants believed they were playing against another person. In another condition they believed they were playing against a computer. Comparison of the conditions showed that activity in the mentalising system was much greater when people thought that they were playing against a person, rather than a computer, even though the behaviour of the opponent was the same in both cases.

A similar effect was seen in a PET study in which people played the game, rock-paper-scissors (Gallagher, Jack, Roepstorff, & Frith, 2002). Here also activity in mPFC was greater when people thought that they were playing against a person, rather than a computer, even though the behaviour of the opponent was the same. This effect is not restricted to real-time competitive interactions. It was also seen in a recent study of action observation (Cross et al., 2012). The action observation network, which roughly corresponds to the mirror system, was activated when observing both human and robot actions. The mentalising network, however, was only activated when participants had a prior belief that the agent performing the actions was human. From these interactive paradigms, just as from the many indirect paradigms, we can conclude that the input to the mentalising system is not pinned down to basic information extracted from the kinematics or other aspects of the stimulus. Instead, the system is activated by a higher order belief about the nature of the stimulus.

A Shock to the System

One of the many snares and delusions associated with brain imaging is the belief that locating the brain regions associated with an ability is all that matters. But location cannot reveal much about the cognitive underpinnings of that ability. We are not fond of *reverse inference* (using brain location to make a claim for a function), as this is a very problematic procedure (Poldrack, 2006). While brain imaging was important to identify the mentalising system, it inevitably led to the question: So what? The next advances in our understanding of mentalising came from experimental studies of behaviour of nonverbal infants and animals.

At around the turn of century, something of a jolt hit our fondly held notions of mentalising. The jolt did not come out of the blue. The first hint was presented in a study by Clements and Perner (1994). In the earlier studies with false belief tasks, preschool children and school children were explicitly asked to say where Maxi would look for his chocolate. Below the age of 4 children typically give the wrong answer, saying that Maxi would look in the blue cupboard where the chocolate actually was. However, in addition to asking this explicit question, Clements and Perner also looked at the children's eye movements to see if they would look at the place where Maxi should go to get his chocolate, given his false belief. Amazingly, at around the age of 3, 90% of the children looked at the correct location (the red cupboard), even though most of them gave verbally the wrong answer (the blue cupboard).

It took more than a decade for similar nonverbal markers of mentalising to be used in studies of younger children. One such measure involves *violation of expectation*. If children expect Maxi to look in the red cupboard, where he originally put the chocolate, they will be surprised and look longer when he goes to the blue cupboard. In a landmark study such behaviour was observed in infants of 15 months (Onishi & Baillargeon, 2005). Soon other studies measured *eye movements* to show that infants will correctly anticipate what an actor will do, even when these actions are determined by the actor's false belief (Southgate, Senju, & Csibra, 2007). A number of studies in different labs and with different paradigms have confirmed these results and tested even younger children. For example, Kovács, Téglás, and Endress (2010) provided striking evidence that human infants as young as 6 months old are able to take account of an observer's belief. In this study eye gaze was measured in infants and reaction time in adults. Both were affected in the same way by observing a scenario where a protagonist had a false belief.

There are now several studies suggesting that great apes can take account of false beliefs (see, for example, Krupenye, Kano, Hirata, Call, & Tomasello, 2016). Rhesus monkeys were observed to steal grapes from a human competitor who was not able to see the grapes, rather than from one who could see them (Flombaum & Santos, 2005). In birds too, evidence was found for the ability to take account of a conspecific's knowledge (e.g. in scrub-jays caching behaviour; Clayton, Dally, & Emery, 2007): re-caching occurred only when the scrub-jay who had seen the first hiding place was no longer present.

Our ideas had to change. Previously, we had occasionally toyed with the idea that mentalising is a uniquely human ability (C. D. Frith & U. Frith, 2007; U. Frith & C. D. Frith, 2010). This was based on the seemingly solid findings that ToM, as assessed by the Maxi type of task, was present roughly from age 4 in human children, but not before. On the face of it, this suggested that it required a certain level of metacognitive ability, unlikely to be present in other animals. Yet, there was also the idea that mentalising is based on an innate cognitive ability (U. Frith, 2013) with evolutionary precursors. This idea was strengthened by the findings of mentalising very early in development and behaviour suggestive of mentalising in other species.

A Reconciliation

The facts that had to be reconciled were these: Human children have the capacity to mentalise from early infancy, yet it is not until age 4 that they can explain why Maxi looks in the wrong place for his chocolate. 'He didn't see that his mother had moved it'. Below that age children cannot reason like this, but they still take account of false beliefs, as evident in their eye gaze. An explanation offered itself, which by hindsight we can now see as part of the Zeitgeist, in the form of dual-process theory (J. St B. T. Evans & Stanovich, 2013) as popularised in Kahneman's account, *Thinking Fast and Slow* (Kahneman, 2011). This is the proposal that there are two kinds of mentalising, implicit and explicit (Apperly & Butterfill, 2009), and that both explicit and implicit processes exist side by side.

In autism research too, a reconciliation was necessary between seemingly contradictory findings. Autistic individuals, especially if they are unaffected by intellectual disability, are able to pass Maxi/Sally-Ann type tasks, although at a later age (Happé, 1995). Here too, the assumption of two kinds of mentalising offered an answer. Explicit mentalising could be achieved via compensatory processes, using a slower route (U. Frith, 2004). Implicit mentalising, however, was proposed to be faulty, presumably due to some disconnection in the mentalising system of the brain.

To test this hypothesis, Senju, Southgate, White, and Frith (2009) selected autistic adults, who were highly competent at solving explicit mentalising tasks, and measured their anticipatory eye gaze, using a paradigm originally developed for infants. As predicted, these participants did not automatically anticipate the location where a protagonist believed a hidden object would be. This was in contrast to the eye gaze of neurotypical adults, who, just like young infants, spontaneously anticipated the expected location. This finding confirmed the hypothesis and has subsequently been replicated with specially designed tasks (Rosenblau, Kliemann, Heekeren, & Dziobek, 2015; D. Schneider, Slaughter, Bayliss, & Dux, 2013; Schuwerk, Jarvers, Vuori, & Sodian, 2016).

How to Study Implicit Mentalising?

It is easy to say that implicit mentalising is an automatic, unintentional and unconscious process. But it is hard to demonstrate it in an appropriate task with adults (D. Schneider, Slaughter, & Dux, 2017). The reason is a possible confusion with the term *implicit:* here it applies to the cognitive process, not to the task. Since adults can use either conscious or unconscious abilities, it is hard to control which are used at the moment of testing. For example, if participants are asked simply to observe the animated triangles, this is an implicit mentalising *task*, as there are no *explicit* instructions to think about mental states. However, if the participants started spontaneously thinking about mental states consciously, then they would be engaged in an explicit *process* of mentalising.

It is notoriously difficult to demonstrate the existence of unconscious processes, and even more difficult to elucidate their neural underpinnings. One of the more robust demonstrations of automatic mentalising involves taking account of the viewpoint of another. Having a different spatial viewpoint can create incongruence of knowledge, since what one person can see often differs from what another person can see. Many studies have demonstrated an effect of such incongruence. For example, if I can see everything in a room, I will take longer to report what a person in the room can see, when this is different from what I can see (see Fig. 1). I need to recognise that she cannot see the pictures on the wall behind her. An egocentric bias towards my own point of view interferes with the task of inferring what another can see (Royzman, Cassidy, & Baron, 2003).

Dana Samson and her colleagues reported a novel twist on this phenomenon (2010), which shows a detrimental effect even when there is no need to represent the other person's viewpoint. Participants were never asked how many pictures the other person could see, but only how many they could see. Nevertheless, the mere presence of another person in the room with a different viewpoint (right-hand panel, Fig. 2) slowed down this egocentric response. The process involved here is unintended since it interferes with what the participant is trying to do. The process is automatic since the result has been shown to be unaffected by cognitive load (Qureshi, Apperly, & Samson, 2010). This observation shows that we cannot help

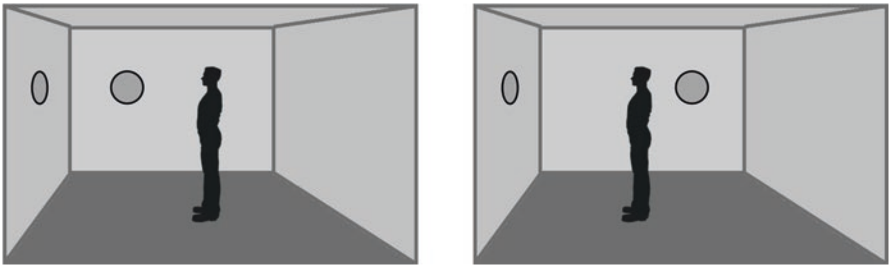


Fig. 1 Left panel, the avatar and the viewer see two disks. Right panel, the avatar sees one disk, while the viewer sees two (after Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010)

taking account of the knowledge of others when it is different from our own. It supports the view that there is a cognitively efficient and automatic process, specialised for tracking own and others' points of view.

The same phenomena can be observed in the understanding of speech. When we hear the sentence '*the girl had a little beak*', we are surprised and mystified by the occurrence of the word '*beak*' (indicated by a large amplitude N400 response in the EEG). This response disappears when we are told that the girl was dressed up as a canary for Halloween. However, the N400 response comes back if we are with a co-listener who does not know this (Jouravlev et al., 2018).

The situation is not so clear for more traditional mentalising tasks involving false beliefs (e.g. Maxi and the chocolate) rather than conflicting viewpoints. Schneider and her colleagues (2017) have shown that the false beliefs of others can be taken account of unintentionally and without awareness. However, they also observed that performance on these tasks was affected by cognitive load. This is evidence of some dependence on domain-general mechanisms such as working memory. Nevertheless, even if implicit, automatic mentalising processes are not necessarily completely encapsulated (Butterfill & Apperly, 2016), they do, however, seem distinctly different from deliberate, explicit mentalising processes which are heavily dependent on general cognitive abilities such as language (Ronald, Viding, Happe, & Plomin, 2006) and executive function (Russell, 1997). Indeed, the very functioning of these general cognitive abilities might be what allows autistic individuals to acquire explicit mentalising.

At this point, the nature of the implicit mentalising process remains obscure. Is it a cognitively efficient but inflexible process, specialised for the tracking belief-like states (Apperly & Butterfill, 2009)? Does it depend on reasoning about behaviour,

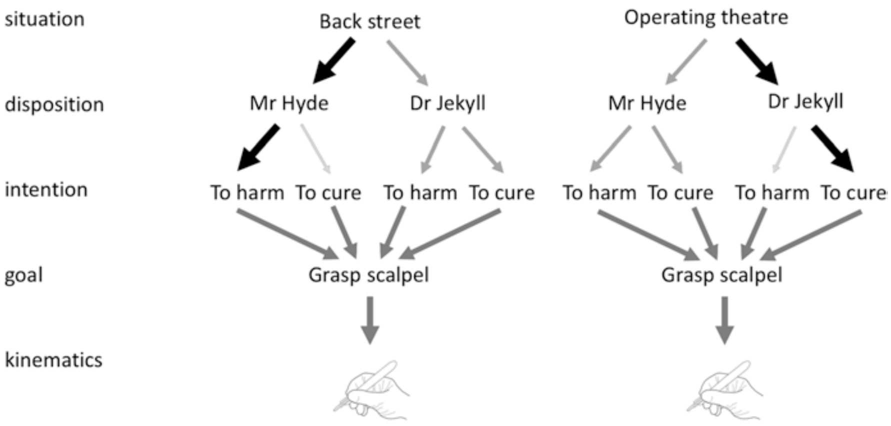


Fig. 2 Our expectations about intentions are constrained by the situation and the disposition of the agent. We expect Dr. Jekyll in the operating theatre to grasp the scalpel in order to cure and predict the kinematic of his movements accordingly. If the kinematics fail to match the prediction then we need to update our estimate of the intention and possibly even of the situation and the disposition (after J. M. Kilner et al., 2007)

rather than about mental states (Perner & Ruffman, 2005)? Or does it conceivably derive from a domain-general process of association learning (Heyes & Frith, 2014)? The situation is no better for explicit mentalising processes. But help is at hand.

Implicit and Explicit Mentalising in the Brain

If there are more than one kind of cognitive process underlying the ability to mentalise, then we would expect this to be reflected in brain activity. Here brain imaging might come to our aid to reveal something about cognition: demonstrating that implicit and explicit mentalising activate different brain regions would be evidence that different cognitive processes are involved.

One problem we already alluded to is that, in many studies, it is unfortunately the *task* rather than the *process* that is implicit. Such studies look at the neural effects of instructing people to mentalise (see, for example, Molenberghs, Johnson, Henry, & Mattingley, 2016). However, as we already pointed out, studying implicit processes is fraught with difficulty, especially in a scanning environment. How can we be sure that no explicit processing occurred? This can be achieved, first by the use of a distracter task and second by using a debriefing task to eliminate participants who reported thinking about mental states (as described in Dana Schneider, Slaughter, Becker, & Dux, 2014). Unfortunately, these requirements make for complicated studies, which can be difficult to interpret.

Schneider et al. (2014) emphasised the lack of overlap between explicit and implicit mentalising processes in terms of brain regions activated. They concluded that left STS and precuneus were key nodes for implicit false belief processing. However, they also observed activity in right TPJ, a major component of the explicit processing network. But this activity was elicited by implicit processing of both false and true beliefs. On the other hand, in a subsequent study, the same group (Naughtin et al., 2017) emphasised the substantial overlap between brain activity associated with implicit and explicit mentalising. Once again STS and precuneus activity was observed for implicit processing of false beliefs, but this time TPJ/pSTS was also strongly implicated. It seems, therefore, that implicit mentalising is associated with activity in a subset of the brain regions that are also associated with explicit mentalising. One area associated with explicit processing that did not seem to be involved in implicit mentalising in these studies was medial prefrontal cortex.

Other Examples of Implicit and Explicit Processes in Social Cognition

Mentalising is not unique among cognitive processes for having an implicit and an explicit form. Indeed, as we have already mentioned, this division fits in with the Zeitgeist and currently prevailing framework for the study of cognition. The distinction between automatic and controlled processes is a feature of cognitive processing in general and of social cognition in particular (J. S. Evans, 2008; Satpute & Lieberman, 2006). With regard to social cognition, the distinction is especially clear in the case of our various prejudices against out-groups (C. D. Frith & U. Frith, 2008).

The presentation of a fear-inducing stimulus can lead to an automatic response such as activation of the amygdala, even when we are unaware of seeing the stimulus (Morris, Ohman, & Dolan, 1999). An implicit form of race prejudice is revealed when people respond to the presentation of a black face with a similar automatic fear response. The amplitude of this response correlates with the degree of prejudice as measured by the Implicit Association Test, but not with explicit (conscious) measures of race prejudice (Phelps et al., 2000). This result shows that there are two forms of race prejudice: implicit and explicit. Within individuals these implicit and explicit forms are relatively independent of each other.

When the faces of black Americans were presented for a longer period (535 ms instead of 30 ms), the amygdala activation was much reduced (Cunningham et al., 2004). Furthermore, the magnitude of activity in the prefrontal cortex predicted how much the amygdala activity would be reduced for the long presentations. This result suggests that activation in dorsolateral prefrontal cortex (DLPFC) and anterior cingulate cortex (ACC) is associated with deliberate attempts to control undesirable prejudicial responses to black faces. The extent of this deliberate control relates to (lack of) explicit race prejudice.

This is an example of what seems to be a general principle for mechanisms of social cognition and, indeed, of all kinds of cognition. At a lower level there are fast, relatively inflexible routines that are largely automatic and may occur without awareness. These routines involve various different brain regions, depending on the sources of the information required. At a higher level there are slow, flexible routines that are explicit and require the expenditure of mental effort. These processes typically involve the frontal cortex.

This framework for the neural basis of social cognition in general and mentalising in particular is nicely summarised by Van Overwalle and Vandekerckhove in their discussion of implicit and explicit social mentalising (2013). ‘Social inferences are hierarchically arranged, with lower-level brain areas (e.g. amygdala) continuously providing valenced information and moderate-level interpretations of behaviours (e.g. TPJ/pSTS) sending information on the agent’s intentions, which feeds higher-level interpretations of the agent in terms of traits’.

We conclude that our two types of mentalising are not completely independent. This is important to bear in mind when trying to understand changes in mentalising behaviour during development and in neurotypical and autistic children.

A Hierarchical System

In this section of our essay we come to the questions and challenges that lie ahead. What are the roles of the various components of the brain's mentalising system? Will computational approaches solve some major questions about the interaction of the putative components? Can we eventually come to understand the relationship between implicit and explicit mentalising?

We are prepared to make some guesses. The comparison of implicit and explicit mentalising discussed above hints at some possible ways of fractionating the system. We can start with the idea that the system forms a hierarchy (see, for example, Gilead, Trope, & Liberman, 2019) and we can relate it to some anatomical facts: The lowest levels of this hierarchy are located in primary sensory and motor cortices while the higher levels, such as prefrontal cortex, are at an increasing synaptic distance from the periphery (Huntenburg, Bazin, & Margulies, 2018). Cognitive processing within this hierarchy involves interactions between bottom-up (feedforward) signals from the periphery and top-down (feedback) signals from the highest level. Explicit processes will be dominated by the top-down signals from the top of the hierarchy, while implicit processes can function in the absence of these top-down signals, as long as any prediction errors can be resolved.

As an example of this approach, we draw on a study of emotional responding related to mentalising from our group where we used a hierarchical model with three levels to explain the results of a brain imaging study (Silani et al., 2008). Participants in this study were shown pictures classified as unpleasant or neutral. The participants performed two different tasks with these pictures. They had to rate the pictures either for the emotion they evoked or for how colourful they were. Unpleasant pictures elicited more activity in the amygdala, whichever rating was being performed. This is consistent with the amygdala showing an automatic response to bottom-up signals of unpleasantness. Comparison of the two rating tasks revealed greater activity in mPFC and precuneus/PCC when participants introspected on their emotional response, whichever type of stimulus was being presented. This is consistent with these being the source of top-down signals required for controlled processing, in this case directing attention to internal states.

In contrast to these regions, the anterior insula seems to occupy an intermediate position in the hierarchy. Activity was seen here when participants were introspecting and an unpleasant stimulus was presented (the task by stimulus interaction). Activity here was also correlated with self-report ratings of empathy (positively) and alexithymia (negatively). This is consistent with the anterior insula providing the neural basis of how we are feeling (Craig, 2002). On the basis of a meta-analysis Gu and colleagues (Gu, Hof, Friston, & Fan, 2013) concluded that the anterior insula integrates bottom-up interoceptive signals with top-down signals to create a state of emotional awareness. So, we might speculate that top-down signals from mPFC, during introspection, enhance the response of the anterior insula to bottom-up signals of unpleasantness coming from the amygdala.

Ochsner, Silvers, and Buhle (2012) tell a similar story in their review of the mechanisms underlying the cognitive control of emotion. They suggest that the downregulation of negative emotion involves a hierarchy of control whereby dorsomedial regions reduce amygdala responses via their impact on ventromedial prefrontal cortex.

This hierarchical account has direct parallels with accounts of the differences in brain activity between conscious (implicit) and unconscious (explicit) processing in visual perception. As long as perceptual processing remains at an unconscious level, activity is restricted to visual processing regions (e.g. fusiform face area for faces and visual word-form area for words). When these stimuli become conscious, additional activity is seen in intraparietal sulcus and DLPFC (faces Beck, Rees, Frith, & Lavie, 2001; words Dehaene et al., 2001). By analogy with the system for visual consciousness, we might expect the roles of DLPFC and IPS to be taken by mPFC and precuneus/PCC in the mentalising system.

We believe there is already some evidence of a hierarchical system for mentalising and here we are willing to place some bets about the role of the three major components of the mentalising system, mPFC, TPJ/pSTS, and precuneus/PCC.

The Role of mPFC

We suggest that mPFC is the source of top-down signals arising at the top of the hierarchy, typical of voluntary and controlled action. In our rather broad-brush approach to the function of mPFC, we will not consider further specialisations within this brain region, such as the distinction between ventral and dorsal mPFC. There is considerable evidence that activity in ventral mPFC is associated with judgements about self and similar others, while activity in dorsal mPFC is associated with judgements about dissimilar others (see, for example, Denny, Kober, Wager, & Ochsner, 2012). We assume that this distinction reflects the problem of prior expectations being less precise when we interact with people who are different from ourselves, indicating that different top-down influences are required.

In a Bayesian framework, these top-down influences concern prior expectations about the situation in which the people find themselves. They might well characterise processes that can become conscious, a fit for explicit mentalising. We might call this the mentalising mode by analogy with the retrieval mode in studies of episodic memory (Lepage, Ghaffar, Nyberg, & Tulving, 2000). This mode is a cognitive state which determines how incoming stimuli will be processed. In terms of brain function, this will be revealed as tonic activity maintained throughout a period in which people are engaged in mentalising.

We have already cited several studies in which this pattern of activity was observed in mPFC. In these studies (Gallagher et al., 2002; Rilling et al., 2004) activity in mPFC was seen when participants believed that they were interacting with a person even though there was no difference in the behaviour of their partner. mPFC activity is also not affected by the precise movements of the agent being observed (Stanley, Gowen, & Miall, 2010). In this study cues derived from prior knowledge were contrasted with stimulus-based cues. The belief, via instruction, that the movements of the agent were based on human movements engaged mPFC more than the belief that the stimuli were computer-generated. In contrast, human-like movements did not engage this region more than less human-like movements.

The suggestion that activity in mPFC represents a tonic state (the mentalising mode) fits with the observation that there is considerable overlap between components of the mentalising system and the so-called default mode network. There is substantial evidence that this cognitive mode is reciprocally suppressed by another cognitive mode (the task-positive mode) that is engaged when people perform various tasks. However, Jack et al. (2013) have shown that there is a set of tasks that does engage the default mode network. These are problems concerned with social cognition (i.e. reasoning about the mental states of other persons). In contrast, the task-positive network is engaged by problems concerned with physical cognition (i.e. reasoning about the casual/mechanical properties of inanimate objects). In this study, where a wide range of tasks was used to characterise the two modes, activity associated with the mentalising mode was most pronounced in mPFC and precuneus/PCC. A similar pattern was observed by Spunt and Lieberman (2012) who asked their participants to consider *why* (mental causation) or *how* (physical causation) an action was performed. The comparison of *why* vs. *how* also revealed activity in mPFC and PCC.

The Role of TPJ/pSTS

In contrast to mPFC and PCC, this region appears to occupy an intermediate position in the hierarchy. Evidence for this position can be found in terms of neural connectivity and cognitive processes.

Hillebrandt, Friston, and Blakemore (2014) used fMRI data from the human connectome project, gained during spontaneous mentalising elicited by the Frith-Happé task of animated triangles. When these movements elicited the perception of intentionality, there was an increase in connectivity between V5, a region concerned with low-level motion detection, and pSTS, a region concerned with biological motion, which is part of the TPJ. This is an example of feedforward connections from the sensory regions into the mentalising system.

Similar results for pSTS, i.e. coupling with visual areas, were obtained by Moessnang et al. (2017) using the same task. They also observed a different pattern of connectivity with mPFC and concluded that this region was involved in metacognitive representation, while pSTS was involved in perception-based processing of social information. Both these studies suggest that TPJ/pSTS is receiving signals from low-level sensory regions.

Klapper, Ramsey, Wigboldus, and Cross (2014), using an imitation interference paradigm, confirm that TPJ occupies an intermediate position in the hierarchy in terms of cognitive processes. They showed that right TPJ was engaged more during an automatic imitation task when both stimulus cues (bottom-up) and knowledge cues (to-down) to human animacy were present compared to when only one or neither cue to human animacy was present.

Observation of intentional actions elicits activity in pSTS/TPJ, particularly if the actions are unexpected. For example, greater activity is seen when people move their eyes in an unexpected direction (away from, rather than towards a stimulus; Pelphrey, Singerman, Allison, & McCarthy, 2003) and also when people appear from behind a barrier at an unexpected time (R. Saxe, Xiao, Kovacs, Perrett, & Kanwisher, 2004). This is further evidence that this region is affected by both prior expectations (top-down) and incoming evidence (bottom-up).

However, this incoming evidence is not restricted to direct visual perception of what people are doing. For example, in the study of Hampton, Bossaerts, and O'Doherty (2008), activity in TPJ/pSTS occurred when the partner in the game made an unexpected choice. In this case the partner's choice was indicated by a symbolic representation. Furthermore, in many of the studies in which TPJ/pSTS is activated, the behaviour of people is described in words.

On the basis of these various studies, we conclude that TPJ/pSTS occupies an intermediate position in a mentalising hierarchy. This region receives evidence (bottom-up) concerning people's behaviour and is influenced by expectations (top-down) about what sort of behaviour would be expected. This sounds suitable for an automatic tracking process and would potentially map onto implicit mentalising.

The Role of Precuneus/PCC (Retrosplenial Cortex)

Here we have to be even more speculative, since the role of this region in mentalising has often gone under the radar. But just because it is obscure does not mean it is unimportant. Even in our first PET study (Fletcher et al., 1995) the precuneus was robustly activated during mentalising. Its function outside mentalising is also obscure since it is engaged by many different tasks. These include spatial navigation and scene processing, episodic memory retrieval, mental imagery and self-referential processing (see, e.g. Chrastil, 2018). We suspect that its role in spatial navigation, for which there is extensive evidence, may provide a clue for its role in mentalising.

In their recent review of the role of this region in spatial navigation, Mitchell, Czajkowski, Zhang, Jeffery, and Nelson (2018) suggest that it has access to the same spatial information represented in different ways (e.g. an egocentric or an allocentric representation of space) and is therefore needed in order to switch between these representations. During navigation this would enable us to establish and maintain our bearings in the scene (Hartley, Maguire, Spiers, & Burgess, 2003). As with TPJ/pSTS, this region does not require direct evidence concerning spatial scenes. It is also engaged by words describing spatial scenes (Auger & Maguire, 2018; Vukovic & Shtyrov, 2017).

Vogeley et al. (2004) found that this region was activated when people had to take the viewpoint of someone else into account (visual perspective taking as in the task shown in Fig. 1). Arora, Schurz, and Perner (2017) compared and contrasted a wide range of brain imaging studies involving visual perspective taking or mentalising. They also concluded that this region has a role in third person perspective taking, which is a common feature of both kinds of task. For example, in a false belief task, it is necessary to represent the protagonist's 'false' view of the state of the world.

Of particular interest in this context is a recent suggestion by Schafer and Schiller (2018). They propose that navigation in the social world involves the same neural mechanisms as navigation in space. This idea had been explored earlier in a study in which participants had to navigate through a social space, interacting with people (avatars) who varied along the dimensions of affiliation and power (Tavares et al., 2015). They found that activity in precuneus/PCC tracked the social distance (e.g. lower affiliation and larger differences in power) between the participant and the avatar they were interacting with. A greater social distance would require a greater shift in point of view in order to take account of the mental state of the interacting partner. This account would suggest that this region is relatively high in the hierarchy, maintaining a point of view, rather than responding to incoming evidence. Possibly this qualifies this region as a staging post joining up implicit and explicit mentalising systems. But this is pure speculation.

Of course, we cannot stop at simply contemplating the role of components as they are all necessarily interacting in the hierarchical system that we believe underpins mentalising. To better understand the nature of the hierarchy and the interactions between the components of the mentalising system, a computational model is needed. Such a model should predict behaviour and also enable better specification of the precise role of the different regions involved using model-based fMRI (O'Doherty, Hampton, & Kim, 2007).

Eventually a full neuroanatomical and evolutionary approach is needed to discover and understand the neural mechanisms of mentalising and their origins. A start has already been made in work on social cognition in primates, which we do not review here. Instead we recommend the comprehensive review by Wittmann et al. (2018).

Computational Approaches to Mentalising

Here is what we consider to be the current frontier of research on the neural basis of mentalising (see, e.g. Rabinowitz et al., 2018). This is where we expect answers in due course to such difficult questions as to the evolutionary basis of mentalising, the development of the cognitive components during individual development, and the differences and similarities between implicit and explicit mentalising.

There are promising approaches based on *decision-making paradigms* in a game format where participants interact with each other and have the chance to adapt their behaviour to changing conditions. Consider, for example, the simple game of hide and seek. Mental states come into play when agents start thinking about their opponents as other agents, 'You think that the tree is a currently good place to hide'. This approach leads to *recursion*, where the next level is, 'you think that I think that the tree is a currently good place to look' and so on (see Crawford, Costa-Gomes, & Iriberri, 2013). Performance can be analysed in terms of the strategies that best account for the behaviour, from simple reinforcement learning to strategies that recursively take mental states into account at ever more sophisticated levels (e.g. Devaine, Hollard, & Daunizeau, 2014; Hampton et al., 2008).

Another promising approach is based on the concept of *prediction errors*. We presume that there must be representations of expected behaviour and movements which can be compared with the observed behaviour. This will generate the *prediction errors* which are used to update the expectations. The expected behaviour will be based on representations of estimated mental states. And at a higher level still there will be top-down constraints on the possible mental goals that might be likely given the context in which the interaction is occurring (reaching for the cup to drink from it or to clean it, situational—Iacoboni et al., 2005) and also what is known about the person being interacted with (reaching a scalpel to cure or to harm, dispositional—Jacob & Jeannerod, 2005).

Kilner et al. (2007) (see also Koster-Hale & Saxe, 2013) have developed an account along these lines to show how predictive coding might be used to infer the causes of an observed action. In this scheme pSTS receives bottom-up input from the visual system about the kinematics of an observed movement. It also receives top-down signals from the mirror system. These signals indicate the kinematics that would be expected on the basis of inferences about the actor's intentions (the movements I would make, if I had that intention). If the observed kinematics do not match (prediction error), then the inference about the intentions must be updated. In this scheme a distinction is made between goals and intentions. The goal is the immediate cause of the movement (the kinematics are determined by the goal of grasping a scalpel). The intention is the cause of the goal (to cure—Dr Jekyll, to harm—Mr Hyde).

As we have already seen, several studies suggest that TPJ/pSTS activity reflects prediction errors. Unexpected actions elicit more activity in this region (Pelphrey et al., 2003; R. Saxe et al., 2004). Studies using model-based fMRI, where there is an explicit search for activity correlated with the prediction errors associated with learning, have also pointed to TPJ/pSTS (Behrens, Hunt, Woolrich, & Rushworth, 2008; Hampton et al., 2008).

But there should also be prediction errors operating throughout the mentalising system hierarchy. This has been explored by Theriault and Young (2017), who used narratives to elicit prediction errors concerning the disposition of the protagonists or norm violations (prescriptive). These prediction errors derived from high-level representations did indeed elicit activity throughout the mentalising system, but the role of the different components of the system was not addressed.

A subsequent study from Thornton, Weaverdyck, and Tamir (2019) looked at the *process of prediction* rather than prediction *errors*. They used narratives to elicit mental state attributions, such as embarrassment and satisfaction, and showed that neural patterns associated with the mental state currently under consideration resembled patterns of likely future states more so than patterns of unlikely future states (i.e. drunkenness is likely to be followed by sleepiness, while embarrassment is unlikely to be followed by satisfaction). This observation suggests that people were automatically predicting what the next mental state was most likely to be. The results also suggest that mental states lie in *a representational space* in which associated mental states are close together. This arrangement can provide the source of prediction errors, as when the following mental state is far away from the previous one.

Thornton and colleagues suggest that mPFC might be involved in *maintaining a model of the other person's mind*, while precuneus/PCC is responsible for *comparing the predictions* made by this model with the observed sequence of events. Earlier in this essay, we mentioned the proposal that mentalising involves navigation through the social world and uses some of the same mechanisms as *spatial navigation* (Schafer & Schiller, 2018). In this context, the precuneus/PCC was characterised as representing differences from others in the social space (Tavares et al., 2015).

Here is a further speculation on the role of mPFC, and there is even some evidence that it might be on the right track. We assume that mPFC is constructing a model of an other person's mind in order to predict what they will do next. But, if this is so, what kind of models would these be? At the most fundamental level the appropriate model will be determined by *what sort of agent* we think we are interacting with. If we think that the agent is a mindless automaton, then mPFC need not come into play at all since domain-general processes such as reinforcement learning will be sufficient to predict what it will do next. mPFC comes into play when we believe that the agent we are interacting is an agent like us and is trying to predict what we are going to do next. A simple agent of this kind would predict what we are going to do on the basis of knowledge of our current preferences (Robalino & Robson, 2012). In other words, it is thinking about *our* mental state. A more sophisticated agent might predict what we are going to do by trying to infer what we know about its preferences. In other words, it is thinking about what we think about *its* mental states. This is an example of recursion (Crawford et al., 2013). We can classify other agents in terms of the depth of recursion they are using when they interact with us.

There are several studies suggesting that mPFC activity relates to considerations about depth of recursion. The study of Hampton et al. (2008) used the 'influence' model to explain the behaviour of the participants playing the inspector game. The influence model involves two levels of recursion since the player represents the opponent's representation of the player's intended action. The more players used this influence model, the greater was the activity observed in mPFC (coordinates −3,51,24). The beauty contest game (Keynes, 1936, Chap. 12) also requires recursion for successful play. In this game players have to choose a number between 1 and 100, with the winning guess being some fraction (e.g. half) of the average guess. The average random guess will be 50. A more sophisticated player takes this into account and guesses 25, an even more sophisticated player guesses 12, and so on.

The average level of recursion in such games is only about 1.5 (Camerer, Ho, & Chong, 2004; Schou, 2005). Coricelli and Nagel (2009) scanned people while playing this game and showed that greater recursion was associated with greater activity in mFPC (coordinates 3,48,24).

These games are both competitive, but recursion can also be needed in cooperative games. An example is the stag hunt game (Skyrms, 2003). The players in this game will maximise their reward if they hunt a stag, but they will only succeed if they cooperate. So, player A must believe that player B will cooperate and must also believe that player B believes that player A will cooperate, and so on. Yoshida, Seymour, Friston, and Dolan (2010) developed a computational model of play in this game with various levels of recursion represented. When participants played with an agent, activity in mPFC (coordinates −6,54,14) reflected the tracking of the depth of recursion used by the agent. The proximity of the peak activity observed in these three studies is striking.

Metacognition

These results are all consistent with the idea that mPFC might be involved in maintaining a model of the other person's mind (Thornton et al., 2019). But there is extensive research suggesting that mPFC is also concerned with knowledge of our own minds as well as those of other people (e.g. Rebecca Saxe, Scholz, Moran, & Gabrieli, 2006). Can the same mechanisms be applied to the self and the other? We believe that the link between self and other can be made through the study of *metacognition* (Beran, Brandl, Perner, & Proust, 2012). Metacognition is often defined as the monitoring and control of cognitive processes and could, in principle, be applied to cognitive processes in the other as well as the self.

This possibility is nicely demonstrated in a recent computational account of the processes underlying metacognition (Fleming & Daw, 2017). Here, a framework is presented in which the mechanism for monitoring our own decisions would apply equally to monitoring the decisions of others. This could present a real advance in understanding why mentalising works to infer our own mental states as well as those of others—without invoking mirroring or 'simulation' (Gallese & Goldman, 1998; Gordon, 1986).

Metacognition, in Fleming and Daw's framework, is a second-order process in which there is a decoupling between decision-making mechanisms and confidence estimation. In other words, an estimate of confidence does not directly emerge from the computations on which the decision is based (first order). Rather, the estimate of confidence involves making inferences about the decision-making process. Just as a decision involves making inferences about the state of the world, so estimating confidence involves making inferences about the state of the decision-maker. As Fleming and Daw point out, there is a symmetry here between evaluating one's own actions and those of another actor. The decoupling between the metacognitive monitor and decision-making mechanism reminds us of Leslie's original formulation of the critical feature in the attribution of mental states, namely decoupling a representation from its reference to the actual state of affairs (U. Frith, Morton, & Leslie, 1991; Leslie, 1987).

Conclusions

Our discussion of computational approaches to mentalising reveals how much progress has been made since our 1999 paper. However, it also reveals how much still needs to be done. While there are several new and exciting clues about the various computational processes that are required, it is not at all clear how they will fit together or how precisely they are instantiated in the brain. For example, how does the second-order model of metacognition relate to estimating to what extent the other is engaged in recursion?

There is also a great need for fresh experimental paradigms for studying mentalising. In looking at the recent literature, we were struck by the continued emphasis on observation rather than participation. Most studies still present participants with descriptions of behaviour and ask hypothetical questions. What people think they would do does not always relate very well to what they actually do (FeldmanHall et al., 2012). We would also like to see more second-person neuroscience (Schilbach et al., 2013) in which participants interact with other agents in real time (see, for example, Sevgi, Diaconescu, Tittgemeyer, & Schilbach, 2016).

The other development we would like to see relates to the current interest in predictive coding. This approach has proved very fruitful for the study of mentalising. The basic idea is that, when a prediction error occurs, we need to update our belief about the mind of the other. In this way we can minimise the prediction errors associated with perception. But there is another way to deal with prediction errors within this framework, known as *active inference* (Friston & Frith, 2015). Rather than changing our beliefs to suit the world, we change the world to suit our beliefs. In the context of social interactions, for example, this could involve changing the mental state of the other via teaching or deception (as opposed to physical coercion). Such studies would provide important new evidence for our understanding of mentalising.

We anticipate that the next 20 years will see as exciting developments in the study of mentalising as the last 20. We look forward to writing another account at that time (you never know!).

Acknowledgements We are grateful to Richard Passingham, Leo Schilbach, Kevin Ochsner, and Michael Gilead for their insightful comments on earlier drafts of this chapter.

Chris and Uta Frith receive no funding and declare no conflicts of interest.

References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953–970.

Arora, A., Schurz, M., & Perner, J. (2017). Systematic comparison of brain imaging meta-analyses of ToM with vPT. *BioMed Research International, 2017*, 6875850. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28367446

Auger, S. D., & Maguire, E. A. (2018). Retrosplenial cortex indexes stability beyond the spatial domain. *Journal of Neuroscience, 38*(6), 1472–1481.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*(1), 37–46.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1986). Mechanical, behavioural and intentional understanding of picture stories in autistic children. *British Journal of Developmental Psychology, 4*(2), 113–125.

Beck, D. M., Rees, G., Frith, C. D., & Lavie, N. (2001). Neural correlates of change detection and change blindness. *Nature Neuroscience, 4*(6), 645–650.

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature, 456*(7219), 245–249.

Bennett, J. (1978). Some remarks about concepts. *Behavioral and Brain Sciences, 1*(4), 557–560.

Beran, M. J., Brandl, J. L., Perner, J., & Proust, J. (2012). *Foundations of metacognition*. Oxford: Oxford University Press.

Bonda, E., Petrides, M., Ostry, D., & Evans, A. (1996). Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *The Journal of Neuroscience, 16*(11), 3737–3744.

Brass, M., & Heyes, C. (2005). Imitation: Is cognitive neuroscience solving the correspondence problem? *Trends in Cognitive Sciences, 9*(10), 489–495. https://doi.org/10.1016/j.tics.2005.08.007. S1364-6613(05)00238-X [pii].

Brentano, F. (1995/1874). *Psychology from an empirical standpoint*. London: Routledge.

Brothers, L. (1990). The social brain: A project for integrating primate behavior and neurophysiology in a new domain. *Concepts in Neuroscience, 1*, 27–51.

Brothers, L., Ring, B., & Kling, A. (1990). Response of neurons in the macaque amygdala to complex social stimuli. *Behavioural Brain Research, 41*(3), 199–213.

Brunet, E., Sarfati, Y., Hardy-Bayle, M. C., & Decety, J. (2000). A PET investigation of the attribution of intentions with a nonverbal task. *NeuroImage, 11*(2), 157–166.

Butterfill, S. A., & Apperly, I. A. (2016). Is goal ascription possible in minimal mindreading? *Psychological Review, 123*(2), 228–233.

Byrne, R., & Whiten, A. (Eds.). (1989). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford: Oxford University Press.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics, 119*(3), 861–898.

Castelli, F., Frith, C., Happe, F., & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain, 125*(Pt 8), 1839–1849.

Chrastil, E. R. (2018). Heterogeneity in human retrosplenial cortex: A review of function and connectivity. *Behavioral Neuroscience, 132*(5), 317–338.

Clayton, N. S., Dally, J. M., & Emery, N. J. (2007). Social cognition by food-caching corvids. The western scrub-jay as a natural psychologist. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 362*(1480), 507–522.

Clements, W. A., & Perner, J. (1994). Implicit understanding of false beliefs. *Cognitive Development, 9*(4), 377–395.

Coricelli, G., & Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 106*(23), 9163–9168.

Craig, A. D. (2002). How do you feel? Interoception: The sense of the physiological condition of the body. *Nature Reviews. Neuroscience, 3*(8), 655–666.

Crawford, V. P., Costa-Gomes, M. A., & Iriberri, N. (2013). Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature, 51*(1), 5–62.

Cross, E. S., Liepelt, R., Hamilton, A. F., Parkinson, J., Ramsey, R., Stadler, W., & Prinz, W. (2012). Robotic movement preferentially engages the action observation network. *Human Brain Mapping, 33*(9), 2238–2254.

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of black and white faces. *Psychological Science, 15*(12), 806–813.

de Lange, F. P., Spronk, M., Willems, R. M., Toni, I., & Bekkering, H. (2008). Complementary systems for understanding action intentions. *Current Biology, 18*(6), 454–457.

Dehaene, S., Naccache, L., Cohen, L., Bihan, D. L., Mangin, J. F., Poline, J. B., & Riviere, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience, 4*(7), 752–758.

Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences, 1*(4), 568–570.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience, 24*(8), 1742–1752.

Devaine, M., Hollard, G., & Daunizeau, J. (2014). The social Bayesian brain: Does mentalizing make a difference when we learn? *PLoS Computational Biology, 10*(12), e1003992.

Evans, J. S. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278.

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241.

Fan, Y.-T., Decety, J., Yang, C.-Y., Liu, J.-L., & Cheng, Y. (2010). Unbroken mirror neurons in autism spectrum disorders. *Journal of Child Psychology and Psychiatry, 51*(9), 981–988.

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition, 123*(3), 434–441.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review, 124*(1), 91–114.

Fletcher, P. C., Happe, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S., & Frith, C. D. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition, 57*(2), 109–128.

Flombaum, J. I., & Santos, L. R. (2005). Rhesus monkeys attribute perceptions to others. *Current Biology, 15*(5), 447–452.

Friston, K. J., & Frith, C. D. (2015). Active inference, communication and hermeneutics. *Cortex, 68*, 129–143.

Frith, C. D., & Frith, U. (1999). Interacting minds--A biological basis. *Science, 286*(5445), 1692–1695.

Frith, C. D., & Frith, U. (2007). Social cognition in humans. *Current Biology, 17*(16), R724–R732.

Frith, C. D., & Frith, U. (2008). Implicit and explicit processes in social cognition. *Neuron, 60*(3), 503–510.

Frith, U. (1989). *Autism: Explaining the enigma*. Oxford: Blackwell.

Frith, U. (2004). Emanuel Miller lecture: Confusions and controversies about Asperger syndrome. *Journal of Child Psychology and Psychiatry, 45*(4), 672–686.

Frith, U. (2013). *Are there innate mechanisms that make us social beings?* Paper presented at the Neurosciences and the Human Person: New Perspectives on Human Activities, Vatican City. Retrieved from http://www.pas.va/content/dam/accademia/pdf/sv121/sv121-frithu.pdf

Frith, U., & Frith, C. (2010). The social brain: Allowing humans to boldly go where no other species has been. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 365*(1537), 165–176.

Frith, U., & Happé, F. (1994). Autism: Beyond "theory of mind". *Cognition, 50*(1), 115–132.

Frith, U., Morton, J., & Leslie, A. M. (1991). The cognitive basis of a biological disorder: Autism. *Trends in Neurosciences, 14*(10), 433–438.

Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage, 16*(3 Pt 1), 814–821.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2*(12), 493–501.

Gilead, M., Trope, Y., & Liberman, N. (2019). Above and beyond the concrete: The diverse representational substrates of the predictive brain. In *Behavioral and brain science* (pp. 1–63). Cambridge: Cambridge University Press.

Goel, V., Grafman, J., Sadato, N., & Hallett, M. (1995). Modeling other minds. *Neuroreport, 6*(13), 1741–1746.

Gordon, R. M. (1986). Folk psychology as simulation. *Mind & Language, 1*(2), 158–171.

Grèzes, J., Frith, C. D., & Passingham, R. E. (2004). Inferring false beliefs from the actions of oneself and others: An fMRI study. *NeuroImage, 21*(2), 744–750.

Gu, X., Hof, P. R., Friston, K. J., & Fan, J. (2013). Anterior insular cortex and emotional awareness. *The Journal of Comparative Neurology, 521*(15), 3371–3388.

Hamilton, A. F., Brindley, R. M., & Frith, U. (2007). Imitation and action understanding in autistic spectrum disorders: How valid is the hypothesis of a deficit in the mirror neuron system? *Neuropsychologia, 45*(8), 1859–1868.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences of the United States of America, 105*(18), 6741–6746.

Happé, F. G. E. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development, 66*(3), 843–855.

Harman, G. (1978). Studying the chimpanzee's theory of mind. *Behavioral and Brain Sciences, 1*(4), 576–577.

Hartley, T., Maguire, E. A., Spiers, H. J., & Burgess, N. (2003). The well-worn route and the path less traveled: Distinct neural bases of route following and wayfinding in humans. *Neuron, 37*(5), 877–888.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57*, 243–249.

Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science, 344*(6190), 1243091.

Hillebrandt, H., Friston, K. J., & Blakemore, S.-J. (2014). Effective connectivity during animacy perception – Dynamic causal modelling of human connectome project data. *Scientific Reports, 4*, 6240. https://doi.org/10.1038/srep06240

Humphrey, N. K. (1976). The social function of intellect. In P. P. G. Bateson & R. A. Hinde (Eds.), *Growing points in ethology* (pp. 303–317). Cambridge: Cambridge University Press.

Huntenburg, J. M., Bazin, P.-L., & Margulies, D. S. (2018). Large-scale gradients in human cortical organization. *Trends in Cognitive Sciences, 22*(1), 21–31.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology, 3*(3), e79.

Jack, A. I., Dawson, A. J., Begany, K. L., Leckie, R. L., Barry, K. P., Ciccia, A. H., & Snyder, A. Z. (2013). fMRI reveals reciprocal inhibition between social and physical cognitive domains. *NeuroImage, 66*, 385–401.

Jacob, P., & Jeannerod, M. (2005). The motor theory of social cognition: A critique. *Trends in Cognitive Sciences, 9*(1), 21–25.

Jellema, T., & Perrett, D. I. (2003). Cells in monkey STS responsive to articulated body motions and consequent static posture: A case of implied motion? *Neuropsychologia, 41*(13), 1728–1737.

Jouravlev, O., Schwartz, R., Ayyash, D., Mineroff, Z., Gibson, E., & Fedorenko, E. (2018). Tracking colisteners' knowledge states during language comprehension. *Psychological Science, 30*(1), 3–19.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Keynes, J. M. (1936). *General theory of employment interest and money*. London: Macmillan.

Kilner, J. M. (2009). Dissociable functional roles of the human action-observation network (Commentary on E. S. Cross et al.). *The European Journal of Neuroscience, 30*(7), 1382–1382.

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing, 8*(3), 159–166.

Klapper, A., Ramsey, R., Wigboldus, D., & Cross, E. S. (2014). The control of automatic imitation based on bottom–up and top–down cues to animacy: Insights from brain and behavior. *Journal of Cognitive Neuroscience, 26*(11), 2503–2513.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron, 79*(5), 836–848.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science, 330*(6012), 1830–1834.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science, 354*(6308), 110.

Lepage, M., Ghaffar, O., Nyberg, L., & Tulving, E. (2000). Prefrontal cortex and episodic memory retrieval mode. *Proceedings of the National Academy of Sciences, 97*(1), 506–511.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind". *Psychological Review, 94*(4), 412–426.

Mars, R. B., Sallet, J., Neubert, F. X., & Rushworth, M. F. (2013). Connectivity profiles reveal the relationship between brain areas for social cognition in human and monkey temporoparietal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 110*(26), 10806–10811.

Marsh, L. E., & Hamilton, A. F. C. (2011). Dissociation of mirroring and mentalising systems in autism. *NeuroImage, 56*(3), 1511–1519.

Mitchell, A. S., Czajkowski, R., Zhang, N., Jeffery, K., & Nelson, A. J. D. (2018). Retrosplenial cortex and its role in spatial cognition. *Brain and Neuroscience Advances, 2*, 1–13.

Moessnang, C., Otto, K., Bilek, E., Schafer, A., Baumeister, S., Hohmann, S., … Meyer-Lindenberg, A. (2017). Differential responses of the dorsomedial prefrontal cortex and right posterior superior temporal sulcus to spontaneous mentalizing. *Human Brain Mapping, 38*(8), 3791–3803.

Molenberghs, P., Johnson, H., Henry, J. D., & Mattingley, J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience and Biobehavioral Reviews, 65*, 276–291.

Morris, J. S., Ohman, A., & Dolan, R. J. (1999). A subcortical pathway to the right amygdala mediating "unseen" fear. *Proceedings of the National Academy of Sciences of the United States of America, 96*(4), 1680–1685.

Naughtin, C. K., Horne, K., Schneider, D., Venini, D., York, A., & Dux, P. E. (2017). Do implicit and explicit belief processing share neural substrates? *Human Brain Mapping, 38*(9), 4760–4772.

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences, 1104*, 35–53.

Ochsner, K. N., Silvers, J. A., & Buhle, J. T. (2012). Functional imaging studies of emotion regulation: A synthetic review and evolving model of the cognitive control of emotion. *Annals of the New York Academy of Sciences, 1251*, E1–E24.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255–258.

Pellegrino, G. d., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research, 91*(1), 176–180.

Pelphrey, K. A., Singerman, J. D., Allison, T., & McCarthy, G. (2003). Brain activation evoked by perception of gaze shifts: The influence of context. *Neuropsychologia, 41*(2), 156–170.

Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science, 308*(5719), 214–216.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience, 12*(5), 729–738.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10*(2), 59–63.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioural and Brain Sciences, 4*, 515–526.

Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *The Journal of Neuroscience, 18*(6), 2188.

Qureshi, A. W., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: Evidence from a dual-task study of adults. *Cognition, 117*(2), 230–236.

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S. M. A., & Botvinick, M. (2018). Machine theory of mind. *arXiv*. Retrieved from https://ui.adsabs.harvard.edu/\#abs/2018arXiv180207740R

Ramachandran, V. S., & Oberman, L. M. (2006). Broken mirrors: A theory of autism. *Scientific American, 295*(5), 62–69.

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews. Neuroscience, 20*, 495. https://doi.org/10.1038/s41583-019-0179-4

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *NeuroImage, 22*(4), 1694–1703.

Robalino, N., & Robson, A. (2012). The economic approach to 'theory of mind'. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 367*(1599), 2224–2233.

Ronald, A., Viding, E., Happe, F., & Plomin, R. (2006). Individual differences in theory of mind ability in middle childhood and links with verbal ability and autistic traits: A twin study. *Social Neuroscience, 1*(3-4), 412–425.

Rosenblau, G., Kliemann, D., Heekeren, H. R., & Dziobek, I. (2015). Approximating implicit and explicit mentalizing with two naturalistic video-based tasks in typical development and autism spectrum disorder. *Journal of Autism and Developmental Disorders, 45*(4), 953–965.

Royzman, E. B., Cassidy, K. W., & Baron, J. (2003). "I know, you know": Epistemic egocentrism in children and adults. *Review of General Psychology, 7*(1), 38–65.

Russell, J. (1997). How executive disorders can bring about an inadequate 'theory of mind'. In *Autism as an executive disorder* (pp. 256–304). New York, NY: Oxford University Press.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology. Human Perception and Performance, 36*(5), 1255–1266.

Satpute, A. B., & Lieberman, M. D. (2006). Integrating automatic and controlled processes into neurocognitive models of social cognition. *Brain Research, 1079*(1), 86–97.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *NeuroImage, 19*(4), 1835–1842.

Saxe, R., Scholz, J., Moran, J. M., & Gabrieli, J. (2006). Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Social Cognitive and Affective Neuroscience, 1*(3), 229–234.

Saxe, R., Xiao, D. K., Kovacs, G., Perrett, D. I., & Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia, 42*(11), 1435–1446.

Schafer, M., & Schiller, D. (2018). Navigating social space. *Neuron, 100*(2), 476–489.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *The Behavioral and Brain Sciences, 36*(4), 393–414.

Schneider, D., Slaughter, V. P., Bayliss, A. P., & Dux, P. E. (2013). A temporally sustained implicit theory of mind deficit in autism spectrum disorders. *Cognition, 129*(2), 410–417.

Schneider, D., Slaughter, V. P., Becker, S. I., & Dux, P. E. (2014). Implicit false-belief processing in the human brain. *NeuroImage, 101*, 268–275.

Schneider, D., Slaughter, V. P., & Dux, P. E. (2017). Current evidence for automatic theory of mind processing in adults. *Cognition, 162*, 27–31.

Schou, A. (2005, September 22). Gæt-et-tal konkurrence afslører at vi er irrationelle. *Politiken*.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews, 42*, 9–34.

Schuwerk, T., Jarvers, I., Vuori, M., & Sodian, B. (2016). Implicit mentalizing persists beyond early childhood and is profoundly impaired in children with autism spectrum condition. *Frontiers in Psychology, 7*, 1696.

Searle, J. R. (1995). *The construction of social reality*. New York, NY: Simon and Schuster.

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in Asperger syndrome. *Science, 325*(5942), 883–885.

Sevgi, M., Diaconescu, A. O., Tittgemeyer, M., & Schilbach, L. (2016). Social Bayes: Using Bayesian modeling to study autistic trait-related differences in social cognition. *Biological Psychiatry, 80*(2), 112–119.

Silani, G., Bird, G., Brindley, R., Singer, T., Frith, C., & Frith, U. (2008). Levels of emotional awareness and autism: An fMRI study. *Social Neuroscience, 3*(2), 97–112.

Skyrms, B. (2003). *The stag hunt and the evolution of social structure*. Cambridge: Cambridge University Press.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18*(7), 587–592.

Spengler, S., von Cramon, D. Y., & Brass, M. (2009). Control of shared representations relies on key processes involved in mental state attribution. *Human Brain Mapping, 30*(11), 3704–3718.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Oxford: Blackwell.

Spunt, R. P., & Lieberman, M. D. (2012). Dissociating modality-specific and supramodal neural systems for action understanding. *The Journal of Neuroscience, 32*(10), 3575–3583.

Stanley, J., Gowen, E., & Miall, R. C. (2010). How instructions modify perception: An fMRI study investigating brain areas involved in attributing human agency. *NeuroImage, 52*(1), 389–400.

Stel, M., van Dijk, E., & Olivier, E. (2009). You want to know the truth? Then don't mimic! *Psychological Science, 20*(6), 693–699.

Tavares, R. M., Mendelsohn, A., Grossman, Y., Williams, C. H., Shapiro, M., Trope, Y., & Schiller, D. (2015). A map for social navigation in the human brain. *Neuron, 87*(1), 231–243.

Theriault, J. E., & Young, L. (2017). Social prediction in the theory of mind network. *PsyArXiv*. Retrieved from https://psyarxiv.com/hvn54/

Thornton, M. A., Weaverdyck, M. E., & Tamir, D. I. (2019). The social brain automatically predicts others' future mental states. *The Journal of Neuroscience, 39*(1), 140–148.

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30*(3), 829–858.

Van Overwalle, F., & Vandekerckhove, M. (2013). Implicit and explicit social mentalizing: Dual processes driven by a shared neural network. *Frontiers in Human Neuroscience, 7*, 560.

Vogeley, K., May, M., Ritzl, A., Falkai, P., Zilles, K., & Fink, G. R. (2004). Neural correlates of first-person perspective as one constituent of human self-consciousness. *Journal of Cognitive Neuroscience, 16*(5), 817–827.

Vukovic, N., & Shtyrov, Y. (2017). Cortical networks for reference-frame processing are shared by language and spatial navigation systems. *NeuroImage, 161*, 120–133.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655–684.

Wheatley, T., Boncz, A., Toni, I., & Stolk, A. (2019). Beyond the isolated brain: The promise and challenge of interacting minds. *Neuron, 103*(2), 186–188.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs - Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103–128.

Wittmann, M. K., Lockwood, P. L., & Rushworth, M. F. S. (2018). Neural mechanisms of social cognition in primates. *Annual Review of Neuroscience, 41*, 99–118. https://doi.org/10.1146/annurev-neuro-080317-061450

Woodruff, G., & Premack, D. (1979). Intentional communication in the chimpanzee: The development of deception. *Cognition, 7*(4), 333–362.

Yoshida, W., Seymour, B., Friston, K. J., & Dolan, R. J. (2010). Neural mechanisms of belief inference during cooperative games. *The Journal of Neuroscience, 30*(32), 10744–10751.

Part II
The Boundaries of Mentalizing

Early Theory of Mind Development: Are Infants Inherently Altercentric?

Charlotte Grosse Wiesmann and Victoria Southgate

Our daily interaction with others crucially relies on our ability to understand what they think or believe. This ability to represent other individual's mental states has been referred to as *mentalizing* or *Theory of Mind*. Crucially, as opposed to others' behavior or emotional expression, their mental states are not observable and therefore, we need to *infer* what they are thinking. This is why this ability has been viewed as a *theory* of the other's mind (Premack & Woodruff, 1978). Imputing unobservable mental states can be used to make predictions about how others will act that go far beyond what we can conclude merely based on their behavior. As human adults, we often explicitly refer to other people's mental states to make sense of their behavior. For example, when asked why a friend is looking around seemingly searching for something, we might answer that she wants to call someone and is looking for her phone, that she thinks it is in her bag, so she is going to search in there, thus referring to her intention and to her belief about the phone's location. Critically, this answer is independent of where the phone actually is. We could have answered the exact same way if, in fact, she had left her phone in the car. Such utterances clearly show that we make inferences about others' mental states and use

The original version of this chapter was revised. The correction to this chapter is available at https://doi.org/10.1007/978-3-030-51890-5_35

C. Grosse Wiesmann (✉)
Minerva Fast Track Research Group Milestones for Early Cognitive Development, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Department of Psychology, Centre for Early Childhood Cognition, Copenhagen University, Copenhagen, Denmark
e-mail: wiesmann@cbs.mpg.de

V. Southgate
Department of Psychology, Centre for Early Childhood Cognition, Copenhagen University, Copenhagen, Denmark
e-mail: victoria.southgate@psy.ku.dk

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_3

these to explain and predict their behavior. Young children, in contrast, cannot give such elaborate reports, and other species obviously cannot either. Without these explanations, however, we cannot be sure whether children really reason about the other's mental states to predict where she is going to search (i.e., *she thinks her phone is in the bag, so she is going to search for it there*), or whether they predict her behavior based on the real state of the world (i.e., *the phone is actually in the bag, so she is going to search for it there*). It is only in cases where an agent's belief about the world differs from reality that these two strategies yield different predictions. This is why, in response to Premack & Woodruff, predicting an individual's actions based on their false beliefs was suggested as the critical test of a Theory of Mind (Bennett, 1978; Dennett, 1978; Harman, 1978). This test is referred to as the *false belief task* and is still considered as the gold standard to test Theory of Mind abilities.

In a typical false belief task, children witness the story of Sally who puts a marble into her basket and then leaves the room (Baron-Cohen, Leslie, & Frith, 1985). In the meantime, her friend Anne takes the marble out of the basket and puts it into her box. Then, Sally comes back and children are asked where she will search for her marble? As adults, we know that Sally is going to search in the basket, because she doesn't know that her marble has been moved to the box. Until the age of 3 years, however, children usually answer that Sally is going to search in the box, that is, where the marble really is. It is only by the age of around 4 years that children start passing these tests, and do so quite consistently across different versions of the test (Wellman, Cross, & Watson, 2001), including false beliefs about the location of an object—as in the Sally and Anne task—(Wimmer & Perner, 1983) as well as about the content of a container (Hogrefe, Wimmer, & Perner, 1986) or the real identity of an object (Flavell, Green, Flavell, Watson, & Campione, 1986; Gopnik & Astington, 1988). This consistent developmental breakthrough on the different traditional false belief tasks and the strong correlation between performances on these tasks has been the basis for the view that it is around the age of 4 years that children begin to understand others' mental states as representations that may or may not correspond to reality (Astington & Gopnik, 1991; Perner, 1991). In this sense, as a representation of mental representations, Theory of Mind has been referred to as meta-representational (Perner, 1991; Pylyshyn, 1978). According to traditional accounts, this process of meta-representing others' mental representations draws heavily on executive and linguistic resources, and such a meta-representational Theory of Mind would thus develop relatively late in preschool-age once these resources have been acquired (Astington & Baird, 2005; Devine & Hughes, 2014; Wellman et al., 2001).

Infant False Belief Tasks

In the past 15 years, however, a set of new infant false belief tasks have fundamentally questioned this traditional view. These studies, which rely on so-called implicit measures like looking-time, anticipatory looking, and spontaneous participation, showed that infants in their second year of life already have correct expectations where a protagonist with a false belief about an object will search for that object (for a review see, e.g., Scott & Baillargeon, 2017). For example, in 2005, Onishi and Baillargeon used a violation of expectation paradigm to show that 15-month-old infants looked longer at an agent who searched for a toy melon in its actual location although she had a false belief about where that object was, than if the agent had witnessed the toy's transfer to the new location and thus had a true belief about its location. These results suggested that the infants had correct expectations of where an agent with a false belief would search, and were surprised when she searched in the actual location instead. In the following years, this conclusion has been bolstered by a large number of novel infant false belief paradigms that extended the original violation of expectation findings to anticipatory looking and interaction-based methods. The anticipatory looking paradigms showed that infants anticipated with their gaze, where an agent with a false belief about an object will search (e.g., Senju, Southgate, Snape, Leonard, & Csibra, 2011; Southgate, Senju, & Csibra, 2007; Surian & Geraci, 2012). In the interaction-based tasks, infants interpreted the actions or communication of an agent with a false belief in accordance with the agent's belief (Buttelmann, Carpenter, & Tomasello, 2009; Southgate, Chevallier, & Csibra, 2010), and warned her or showed suspense in anticipation of her erroneous search (Knudsen & Liszkowski, 2012b; Moll, Khalulyan, & Moffett, 2017). For example, in a study by Southgate et al. (2010), 17-month-olds approached the box that the experimenter had pointed to in case she knew what this box contained, but went to the other box in case she had a false belief on which toy was in which box (see also Király, Oláh, Csibra, & Kovács, 2018). To date, there are more than 30 studies supporting the view that, by their second year of life, infants take into account other individuals' false beliefs in their expectations and interpretation of the other's actions. These findings have caused a radical overhaul of the traditional view that Theory of Mind develops around the age of 4 years, and triggered one of the most controversial debates in current developmental psychology. Do preverbal infants already have a Theory of Mind? Or how else do they solve these new false belief tasks? And if they do, why then do children consistently fail the traditional explicit ToM tasks until several years later in human development? This debate has additionally been fueled by the non-replication of some infant false belief paradigms (e.g., Dörrenberg, Rakoczy, & Liszkowski, 2018; Grosse Wiesmann, Friederici, Disla, Steinbeis, & Singer, 2018; Kulke & Rakoczy, 2018; Powell, Hobbs, Bardis, Carey, & Saxe, 2018) that by some authors have suggested to point to specific limitations of the infant Theory of Mind tasks (e.g., Grosse Wiesmann, Friederici, Disla, Steinbeis, & Singer, 2018; Powell et al., 2018). Together with the large body of positive findings on Theory of Mind in infancy (see e.g., Kulke & Rakoczy, 2018; Scott & Baillargeon, 2017), these findings call for a parsimonious account of infants' success in many non-verbal or spontaneous false belief tasks before 2 years of age, possible limitations of these abilities, and preschoolers failure on the traditional verbal false belief tasks until around 4 years of age.

Current Theories on Infant False Belief Understanding

Based on the infant false belief findings, some researchers have argued that a meta-representational Theory of Mind is present from very early in infancy (e.g., Scott & Baillargeon, 2017), and may even be innate (Kovács, Téglás, & Endress, 2010; Leslie, 2005). According to these views, it is merely due to extrinsic task-demands on linguistic and executive functions (Scott & Baillargeon, 2017) or pragmatic misunderstandings of the test question (Helming, Strickland, & Jacob, 2014) that young children fail the traditional verbal false belief tasks. Baillargeon, Scott, and He (2010), for example, have argued that it is only the explicit question in the traditional tasks which makes executive demands by requiring children to inhibit their prepotent response to act based on their knowledge of the real location of the object and to select the correct response to pass the test. These arguments mainly focus on the task-related demands of the false belief tasks (e.g., select the correct response, and inhibit the other response in the test situation). What this account ignores are the executive demands that the Theory of Mind process itself makes. For example, reasoning about someone else's beliefs as independent of the real world arguably requires one to encode and store two different representations of the world, to select the appropriate representation, and inhibit the other. This challenge has been argued to cause young preschooler's failure on the traditional verbal false belief tasks (e.g., Carlson, Claxton, & Moses, 2015; Devine & Hughes, 2014; Grosse Wiesmann, Friederici, Singer, & Steinbeis, 2017). The nativist or early infancy accounts of Theory of Mind therefore raise the question how infants are supposed to master these processes with their notoriously immature executive and cognitive resources (Southgate, 2013).

In opposition to early or nativist views of full-fledged Theory of Mind, other researchers have defended the traditional view that Theory of Mind only develops around 4 years, and have argued that infants solve the novel false belief tasks with different processes. In views that discount the involvement of any form of mentalizing, success in the infant false belief tasks is explained by domain-general processes, such as attention or learning statistical associations (e.g., Heyes, 2014; Ruffman, 2014), by learning behavioral rules (Perner & Ruffman, 2005), or by confounds in the original tasks (Kulke, von Duhn, Schneider, & Rakoczy, 2018; J. Phillips et al., 2015). Despite certain limitations on some of the original tasks (Dörrenberg et al., 2018; Grosse Wiesmann et al., 2018; Kulke et al., 2018; Powell et al., 2018), the consistency of the infant findings across a multitude of different tasks and methods casts doubt on these lower-level explanations. An alternative position suggests two different systems that both serve to read other people's minds (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013; Perner & Roessler, 2012), in line with the tradition of dual-process theories in other cognitive domains (e.g., Evans & Stanovich, 2013)—an earlier-developing cognitively less demanding automatic and possibly implicit system and a later-developing cognitively demanding but more flexible system. While the earlier-developing system would allow infants to make correct predictions in some false belief situations, only the later-developing process would allow for fully flexible verbal attribution of false beliefs in any situation. There is some empirical support for different processes underlying the infant anticipatory looking false belief tasks and the traditional verbal false belief tasks—there is a dissociation of performance in these two task types in preschool children on the behavioral level (Grosse Wiesmann, Friederici, et al., 2017), in the brain (Grosse Wiesmann, Friederici, Steinbeis, & Singer, 2020; Grosse Wiesmann, Schreiber, et al., 2017) and in autism (Senju, Southgate, White, & Frith, 2009). In most two-system accounts of Theory of Mind, the two systems are assumed to be independent of each other, and no mechanisms are described for the development from system 1 to system 2. The later-developing system is usually assumed to be the flexible attribution of mental states as representations (our mature meta-representational Theory of Mind), whereas the nature of the early-developing process remains debated. In particular, it is unclear what makes the early process less demanding, and how infants with their limited cognitive resources could master reasoning about false beliefs in these novel tasks.

Demands of Meta-Representational False Belief Reasoning

In the traditional understanding, flexibly reasoning about someone's false belief involves a number of demanding cognitive and executive processes. (1) Children need to be able to represent two divergent representations of the world, their own (or reality) and the other's perspective (to allow them to select from these perspectives appropriately). (2) They need to have a mechanism that allows them to decide which representation to retrieve in which context, i.e., the correct representation to predict physical events in the real world on the one hand (e.g., *the marble is in the box*) and someone's actions on the other hand (e.g., *Sally thinks the marble is in her basket and will therefore search for it there*). (3) They need to activate this correct representation and inhibit the other (e.g., inhibit their own perspective in order to make a correct action prediction). The mechanism that is usually held to enable these steps (or at least step 1 and 2) is a meta-representational format, i.e., forming a representation of a representation. Meta-representation is a hierarchical structure that entails a representation of the world (e.g., the marble is in Anne's box), which includes a representation of others' mental representation (e.g., Sally thinks the marble is in her basket). This structure allows us to represent two representations simultaneously by embedding one into the other and to retrieve our own representation as a primary representation of the world and another person's representation when reasoning about this person within our own primary representation of the world. The traditional false belief tasks test whether children are capable of all three steps of flexible false belief reasoning by asking children to predict someone else's behavior (*Where will Sally search for her marble?*) and, at the same time, asking them about the real state of the world in a memory control question (*And where is the marble really?*). In order to answer both of these questions correctly, children need to encode and store both representations throughout the same situation and, depending on the question, need to retrieve the correct one.

The Altercentric Account

The novel infant false belief tasks, in contrast, do not ask such direct test questions. Instead, they observe children's looking or spontaneous behavior in a situation where an agent has a false belief, and conclude whether the child has a correct expectation of where the agent is going to search. Crucially, these tasks don't have a reality control question to check whether the child correctly represents where the object actually is (Perner, 2011). Instead, the infant false belief tasks usually have a control condition in which the agent has a true belief about the state of the world. This checks whether children make different predictions depending on the agent's belief state, but does not include any check of the child's own representation of the world. One possibility for infants to solve these tasks without the demanding processes of flexible false belief reasoning, therefore, is to *primarily* represent the other's perspective and, in effect, abandon their own perspective (Grosse Wiesmann, 2017; Southgate, 2013, 2020). By fully adopting the other's perspective rather than meta-representing it, infants would be able to represent the other's view without the need to encode two divergent perspectives, store them in a complex hierarchical structure, and inhibit one in favor of the other (Grosse Wiesmann, 2017). This means that rather than meta-representing that Sally *believes* the marble to be in her basket, the children could themselves, in fact, represent the marble as being in the basket.

Why should infants prioritize the altercentric perspective over their own egocentric one? In the first years of life, infants are strongly dependent on other individuals. They rely on others to feed them, to satisfy all their needs, to provide them with information, and to learn from them. Because their own ability to act on the world is very limited, the actions of other agents in their environment will often be more relevant for them than the physical state of the world. During infancy, it could be an adaptive strategy to direct limited attentional resources to the altercentric perspective rather than their own representation of events in the world (Southgate, 2020). Especially, in situations where the other's actions and communication are likely to be relevant, encoding the situation from the other's perspective would be an efficient way to predict and understand their behavior. This would allow young children with limited cognitive and executive resources to generate accurate action predictions and interpret others' actions and communication in accordance with the other's (in most cases accurate) beliefs. This account proposes that, despite not having a full meta-representational Theory of Mind with flexible access to both perspectives, infants from very early in life have a tendency to see the world through the eyes of others, and are thus able to predict their behavior, even in situations where the other's belief may be false.

Review of Empirical Findings in Light of the Altercentric Hypothesis

In the seminal violation of expectation task by Onishi and Baillargeon (2005), for example, seeing how the agent witnesses a toy moving into the yellow box could lead infants to encode the agent's perspective (the toy is in the yellow box) and maintain this perspective despite the fact that they later see that the toy moves to the green box, an event that would not be encoded with the same representational strength because it is not co-witnessed by another agent. As a consequence, infants would correctly predict the agent to search in the yellow box in line with the agent's beliefs, although this would not entail that the infant represents this *as* the other's belief. The altercentric account thus predicts infants' surprise when an agent acts in a belief-incongruent way, explaining data from violation of expectation false belief tasks. In a similar way, encoding the agent's, rather than their own, perspective predicts infants' correct anticipation in previous anticipatory looking studies (e.g., Southgate et al., 2007) and their correct interpretation of others' communication and action prediction in interaction-based studies (e.g., Knudsen & Liszkowski, 2012a; Southgate et al., 2010).

In addition to explaining correct action predictions and interpretation, however, the altercentric account also makes a novel and quite striking prediction: the child itself should also expect the toy to be in the yellow box and not in the green box, where it really is. We might therefore expect infants to be surprised if the toy was revealed in its real location, and would expect them to search for it in the other location in line with the other's perspective. These striking empirical predictions can easily be tested with standard infant methods, such as violation of expectation, anticipatory looking, and spontaneous searching behavior.

There is some empirical support for an altercentric modulation of infants' own expectations (Kovács et al., 2010) and searching behavior (Kampis et al., submitted). In a violation of expectation paradigm, Kovács et al. (2010) showed that infants looked longer at an occluder that revealed nothing when a bystander falsely believed a ball to be behind that occluder than if the bystander knew that, indeed, the ball had left the scene. Importantly, this was observed although the infant themself had witnessed the ball leaving the scene. Infants' expectations of an outcome in the real world was thus modulated by the perspective of the bystander albeit this altercentric perspective was entirely irrelevant to the outcome. Similarly, in a study by Kampis et al. (submitted) infants were shown to search longer in an occluded box when the experimenter falsely believed an object to be inside the box than when the experimenter knew that nothing was in the box. These studies show that infants' expectations of, and search for, objects in the real world were influenced by another person's beliefs about the presence or absence of these objects. What remains open, however, is to what extent infants maintained their own representation of the world. That is, do infants represent both perspectives and their search and looking behavior is merely modulated by the agent's perspective? Or do they indeed primarily adopt the agent's perspective, giving up their own, as predicted by the altercentric account? If infants indeed adopt the altercentric perspective, children's looking and searching behavior should be congruent with the agent's beliefs, but not with their own. This means, in the setting of the previous studies (Kampis et al., submitted; Kovács et al., 2010), they should look and search longer if the agent believes the object to be present than if they themselves believe it to be there—a comparison that, to date, has not been examined. Similarly, in a situation with two locations, under this account, children should expect the object to be in the location where the agent falsely believes it to be, and should be surprised to find it in its real location. The altercentric account thus makes specific predictions that can easily be tested by future research.

What Determines Which Perspective Is Encoded?

Importantly, the altercentric account does not claim that infants cannot encode and remember events based on their own perceptual access. There is ample evidence for infants' object representation and memory, from very early in life (Baillargeon, 1986; Baillargeon, Devos, & Graber, 1989; Wilcox, Nadel, & Rosser, 1996). In situations where an agent with a diverging perspective is present, altercentric biases in the infants' own expectations and searching behavior (Kampis et al., submitted; Kovács et al., 2010) suggest that their representation is modulated by the agent's perception of the situation. The altercentric account explains this modulation by proposing that infants do not form a second *competing* representation of the situation, but instead, preferentially encode those events and changes that are witnessed in common with another agent. This suggestion is based on the idea that an event encoded in the presence of another agent will generate a stronger representation than an event encoded alone, and that the infants' representation will therefore only be updated when a change occurs that is witnessed together with the agent. In contrast, if a change occurs in the absence of the agent, infants will not update the representation that they encoded together with the agent. In the case of a single agent with a false belief, this mechanism would allow infants to make correct action predictions in line with the agent's belief without having to co-represent two conflicting perspectives.

There are, however, obvious restrictions to belief-based action prediction based on this altercentric mechanism. In the case of multiple agents, for example, another agent (that is sufficiently salient to the child) might equally enhance the infant's encoding of an event witnessed in common, so that the infant would end up adopting the perspective of the last agent on the scene. Infants might therefore overgeneralize beliefs across agents and expect a first agent to search in the location where a second agent believes the object to be, thus creating action prediction errors in case a previous agent returns to the scene. There is some evidence that infants overgeneralize preferences from one agent to another and predict a second agent's actions based on the preferences of a first agent (Kampis, Somogyi, Itakura, & Király, 2013), our account predicts that these overgeneralizations would also occur for mental states and false belief scenarios. Furthermore, if a second agent moves the object from the first to the second location as in most typical false belief tasks, infants might well encode the final location better than if no agent had moved the object, and expect the protagonist to search in the last location, in line with children's gaze patterns in recent replication attempts of the original anticipatory looking studies (Dörrenberg et al., 2018; Grosse Wiesmann et al., 2018; Grosse Wiesmann, Friederici, et al., 2017; Kulke et al., 2018), but such a prediction would need to be confirmed with systematic manipulations. A question that arises from this account is what determines whether the presence of an agent will lead to enhanced encoding of an event by the infant? Here we are not suggesting that any agent that might be present in the background of a scene will lead to altercentric encoding of the scene. Instead, it is likely that the presence of an agent will only have an impact on the infant's encoding of the scene if this agent is sufficiently salient, for example, if the agent's actions or communication are likely to be relevant to the infant. Factors like the relationship of the agent to the infant, the saliency of the agent and their perceptual access, their interaction with the goal object, and communicative signals towards the infant should therefore influence whether the infant preferentially encodes the agent's perspective or not. This predicts a certain fragility of altercentric perspective taking, as indicated by recent replication attempts (Dörrenberg et al., 2018; Grosse Wiesmann et al., 2018; Powell et al., 2018) that have shown that young children make correct belief-based action prediction in some test situations, but not in others. Moreover, this account explains findings from systematic manipulations of the test situation that show that emphasizing the agent's perspective rather than the object or child's individual perspective increases 2- to 3-year-old's false belief performance (He, Bolz, & Baillargeon, 2012; Helming et al., 2014; Rubio-Fernández & Geurts, 2013, 2016).[1]

In both traditional and infant false belief task settings, 2–3-year-old children's response depended on how easy the narrative was made for the child to follow the agent's actions and perspective (for a review of these findings, see Helming et al., 2014). For example, Rubio-Fernández and Geurts (2013) showed that 3-year-olds adopted the agent's perspective more easily, if the agent turned away but remained in the scene (thus increasing the salience of the agent's perspective) than if she was absent while the object was moved and she acquired a false belief. In addition, when the object was mentioned in the test question (and thus the child's attention was drawn to the object rather than the agent), children started to make egocentric errors, that is, answered according to their own knowledge about the object's location. In contrast, when the object was not mentioned (i.e., the test question was *where is she going to search?*) children answered correctly according to the agent's perspective. Similarly, in an infant anticipatory looking paradigm, He et al. (2012) showed that 2.5-year-olds anticipated correctly where an agent with a false belief would search, if they were not directly addressed with a test question, but the experimenter spoke to herself (*I wonder where she is going to search*), allowing the children to focus on the agent rather than their own perspective. In contrast, if the experimenter directly addressed the child with the same question while holding direct gaze contact,

[1] In this respect, a recent account suggested by Perner (2016) makes similar behavioral predictions. In contrast to the altercentric account suggested here, in Perner's account, young children are suggested to form two conflicting representations of an object (referred to as *mental files* in his framework)—their own representation (a *regular file*) and the others representation (a *vicarious file*)—but are not able to link these two representations to each other, and therefore rely on external cues in order to choose which representation to retrieve in a given situation. While the altercentric account suggests a predominance of the altercentric perspective in infancy, in Perner's account, it is less clear what determines the dominance of one representation over the other.

children started looking towards the actual object location and thus failed to anticipate the agent's actions to the other, believed location. These experiments indicate that young children's perspective taking is fragile and that, depending on which perspective is highlighted by the narrative, children either correctly use the agent's perspective or falsely use their own to generate action predictions. The findings suggest that, from at least 2 years of age, children are not uniquely altercentric, but that they might adopt an egocentric or an altercentric perspective depending on the saliency and relevance of different aspects of the narrative. This is also supported by the findings that children from around 2 years of age adapt their communication depending on the knowledge or ignorance of their communication partner and themselves (e.g., Goupil, Romand-Monnier, & Kouider, 2016; Knudsen & Liszkowski, 2012a, 2012b; Koenig & Harris, 2005) suggesting that they have some insight into others' or their own knowledge states. Knowledge or ignorance, however, does not require children to represent two conflicting representations (e.g., Hogrefe et al., 1986).

Why Do 3-Year-Olds Make Egocentric Errors?

Moreover, in the traditional false belief tasks, 3-year-olds typically make egocentric errors and answer in accord with their own knowledge when asked about the agent's perspective. This error shows that the egocentric perspective is clearly available to 3-year-olds, and, by this age, seems to be more salient than the altercentric perspective. There are a number of factors that are likely to contribute to the increasing relevance of the self-perspective over the second and third year of life. Maturing motor abilities throughout the first years of life allow the infant to develop more and more possibilities to act on the world. As a consequence, the infants' own actions gain increasing importance compared to those of other agents. It is not before infants acquire a concept of themselves as subject of their own actions and experiences that they can attribute their own perceptual experiences to themselves as agents in the world.

Despite having a schema of their own body from the first months of life and a rudimentary sense of agency (e.g., Rochat, 2010; Rochat & Striano, 1999; Rovee-collier, 1978), infants are not believed to develop an explicit concept of themselves as subject of their own first-person experience and actions before the middle of their second year of life when infants start recognizing themselves in the mirror (e.g., Amsterdam, 1972; Musholt, 2012; Rochat, 2010). Mirror self-recognition, in particular in relation to self-conscious emotions such as shame or embarrassment, shown to occur from around 2 years of age, has been argued to reflect an understanding of the self as seen by others (e.g., Rochat, 2010). Around the same age and in relation to their mirror self-recognition, infants acquire an explicit verbal referent for their self-concept and start referring to themselves as *I* or *me* (Lewis & Ramsay, 2004). It is arguably only when an explicit self-concept is available, that it is possible to reference one's own experiences and actions to this concept of the self as the subject of these experiences.

Therefore, although infants seem to encode events based on their perceptual experiences long before the age of 2 years (e.g., Baillargeon, 1986; Baillargeon et al., 1989), it is only once they have a self-concept that they can attribute these experiences to this self. In this sense, the emergence of a self-concept highlights the self as the subject, and possible referent, of experiences and actions. Now, it is no longer only the perception of others that enhances the encoding of events, but these events can also be referenced to the self, creating a mechanism for enhanced encoding of the self-perspective. This leads to a competing egocentric perspective and might thus produce errors in false belief tasks typically observed in 3-year-olds. Indeed, it has been shown that self-referencing (that is, encoding events in relation to the self) enhances the memory for these events in adults as well as in children from at least 3 years of age (Cunningham, Brebner, Quinn, & Turk, 2014; Ross, Hutchison, & Cunningham, 2020; Symons, Johnson, Symons, & Johnson, 1997). This self-reference effect is not likely to be possible in the absence of an explicit concept of the self. With the emergence of an explicit self-concept, self-referencing might thus yield a new mechanism of preferential encoding of events witnessed by the child themself, which starts to compete with the suggested altercentric encoding mechanism. These two mechanisms might lead to competing representations—an egocentric representation informed by events perceived by the child and an altercentric representation informed by events perceived by the other. We suggest that it may be the competition of these two representations that occurs as infants develop self-representation, which finally leads infants to become aware of the existence of two perspectives. These might then become simultaneously accessible without necessarily being meta-represented yet. However, once infants become aware of this conflict, they may become motivated to reconcile it. This may be one factor in the realization that this conflict can be resolved by attributing one of these representations to the other person, for example, as a belief that may be false.

Indeed, there is empirical evidence that from the age of about 2 years, both the ego- and altercentric perspective seem to be available within the same false belief setting (Buttelmann et al., 2009; He et al., 2012; Helming et al., 2014; Moll et al., 2017; Moll, Kane, & Mcgowan, 2016). Two- and 3-year-olds predicted where an agent with a false belief was going to search for an object either correctly in accordance with the agent's false belief or incorrectly based on their own knowledge, depending on small differences during the test question. For example, children made egocentric errors if they were directly addressed with a test question keeping direct gaze, but anticipated correctly if the experimenter asked the very same question (*I wonder where she is going to look*) to herself while looking up like in her own thoughts (He et al., 2012). Furthermore, 3-year-olds answered correctly where an agent was going to search or not, depending on whether the searched object was mentioned in the test question or not (i.e., *Where is she going to search (for her marble)?*), and answered correctly if this test question was asked first, but incorrectly if it was asked after the reality control question that emphasized the child's view on the situation (Rubio-Fernández & Geurts, 2013, 2016). This shows that at the time point of the test questions both perspectives seem to be available to the child, and the child retrieved one or the other representation depending on contextual cues given in the test question. If the test questions emphasized the agent's actions (e.g., by passive viewing of the agent's actions without direct test question addressed to the child), children answered in accordance with the agent's perspectives. In contrast, if the test question emphasized the child's view by addressing the child with a direct test question or focusing on the object, the child answered in accordance with its own perspective. Moreover, even within the same task, toddlers were able to interpret a communication by the agent in line with the agent's perspective, but then act on the object in accordance with their own view (Buttelmann et al., 2009). Finally, 2.5- and 3-year-olds showed increased suspense when watching a scene in which an agent had a false belief about an outcome (e.g., expected a cookie tin to be full although it was almost empty, Moll et al., 2016, 2017), but not when the agent knew about the outcome, suggesting that toddlers co-represented both perspectives and anticipated the agent to be disappointed about the outcome (e.g., the almost empty cookie tin).

All these findings indicate that there are two conflicting representations available to the child from at least 2 years of age, but that until the age of around 4 years, children fail to flexibly select the correct perspective in any given context. In direct test situations, the egocentric perspective is predominant and thus produces the typical egocentric errors observed in the traditional false belief tasks. In cooperative settings (e.g., when the agent and child have the same intention, Buttelmann et al., 2009; Helming et al., 2014) or when the child's perspective stays in the background (e.g., when passively observing a scene, as in He et al., 2012; Moll et al., 2016, 2017), the agent's actions and perspective are at the center of attention, priming the child towards the altercentric representation and enabling the child to reason in line with the agent's false belief.[2]

What Makes 4-Year-Old Children Master the Traditional False Tasks?

The above findings of context-dependent belief reasoning in 2- to 3-year-olds suggest that both perspectives are available to the child—an egocentric and an altercentric representation—and can be used to predict and interpret others' actions. If young children represent both perspectives—their own and the others—why then do they fail to reason about others' false beliefs in the traditional false belief tasks until the age of 4 years?

[2] While other authors have suggested that children younger than 4 years of age might not be able to represent two conflicting perspectives (Phillips & Norby, 2019), these results have also been discussed within accounts by Helming et al. (2014) and Perner (2016) who offer a different theoretical explanation for the findings. Both accounts, however, fail to offer an explanation for the altercentric perspective observed in infancy (Kampis et al., under revision; Kovács et al., 2010; Scott & Baillargeon, 2017) and the change to an apparently predominant egocentric perspective in 3-year-olds.

As argued above, in addition to representing two diverging perspectives, fully flexible false belief reasoning makes a number of additional complex cognitive demands. (1) Children need to have a mechanism to decide which perspective to use in which context; (2) Even if children know which perspective is the correct one in a given context, retrieving this perspective and inhibiting the other makes non-trivial demands on executive functions. The empirical findings indicate the availability of an ego- and altercentric representation in 2- and 3-year-olds, but their failure to flexibly reason about false beliefs in the traditional direct verbal tasks. This suggests that it is one of the above capacities that children lack before the age of 4 years.

The required mechanism needs to allow children to distinguish accurate representations of the real world from mental representations that someone holds about the world. They need to understand that while direct representations of the world are based on perceptual input, mental representations are independent of the real world and may or may not be accurate representations of it. The mechanism needs to faultlessly disclose that an agent will always act based on his or her mental representation of the world, whereas objects and events in the world are described by the representation that the child formed based on perceptual input.

A mechanism that allows for these distinctions is a meta-representation. In a meta-representational structure, our understanding of what Sally thinks about the marble is not a direct representation of the marble based on perceptual input, but instead our representation of Sally's mental representation (i.e., what we think what Sally thinks about the marble), and it is therefore independent of the perceptual world. This structure also allows us to understand our own representations of the world as a mental representation that can be accurate or false, and thus reflect about our own false beliefs—a thought that would not be possible with only two competing representations, an egocentric and an altercentric one. Indeed, it is around the age of 4 years that children start attributing false beliefs to themselves (Hogrefe et al., 1986), at the same time when they stop making egocentric errors when directly asked to attribute false beliefs to others in the traditional false belief tasks (Wellman et al., 2001).

These findings suggest that, despite the availability of two diverging perspectives, 3-year-old children either lack a meta-representational structure that would allow them to reliably identify the correct representation in a given situation, or the executive functions to handle such a structure.

Indeed, handling a meta-representational structure makes a number of complex demands that are likely to heavily draw on children's general cognitive and executive resources. It requires children to embed one representation into another in a hierarchical way, to activate the correct level of the representation and inhibit the others. Such cognitive and executive abilities are notoriously poor in infants and toddlers (Diamond, 2012), show a steep development over preschool-age, and predict children's emerging success in the traditional false belief tasks (e.g., Devine & Hughes, 2014). Passing these tasks (as opposed to earlier success in the implicit false belief tasks) is strongly related with executive functions (in particular, embedded conditional reasoning and conflict inhibition) and hierarchical processing abilities (Carlson et al., 2015; Devine & Hughes, 2014; Grosse Wiesmann, Friederici, et al., 2017), and is related to the maturation of brain regions involved in inhibition and hierarchical processing (Grosse Wiesmann, Schreiber, et al., 2017). The breakthrough in the traditional false belief tasks—marking the emergence of fully flexible, context-independent false belief reasoning—is thus indeed related to developments in these domains, which might enable either forming, or handling, a meta-representation.

For infants and toddlers with their notoriously low executive functions, the suggested altercentric bias might therefore be an effective "mentalizing" mechanism, which in most cases generates correct predictions, while avoiding the complex cognitive demands of meta-representation.

Finally, even in adults, the suggested altercentric mechanism of preferential encoding events that are witnessed by a relevant agent might be an efficient mentalizing strategy when little cognitive resources are available. In adults, who like preschoolers, have an explicit self-concept, this strategy would result in competing ego- and altercentric representations that are activated depending on the relevance of the agent's or one's own perspective. This would explain why, under time pressure or in cognitively demanding tasks, even adults suffer from altercentric biases in their own judgement of a situation, or from egocentric biases when asked to judge someone else's perspective (Kovács et al., 2010; Samson, Apperly, Braithwaite, Andrews, & Scott, 2010; Sommerville, Bernstein, & Meltzoff, 2013; Surtees, Apperly, & Samson, 2016; Van Der Wel, Sebanz, & Knoblich, 2014).

Conclusions

The past 1.5 decades of Theory of Mind research has been puzzled by the apparently incompatible findings that infants younger than 2 years of age pass a variety of newly developed implicit false belief tasks, but consistently fail traditional false belief tasks until the age of 4 years. Here, we offer a theoretical framework that explains infants' and young preschoolers' success on implicit false belief tasks, the fragility of this success in early preschool-age (e.g., Grosse Wiesmann et al., 2018; He et al., 2012; Rubio-Fernández & Geurts, 2013; Setoh, Scott, & Baillargeon, 2016), 3-year-olds' egocentric errors in the traditional false belief tasks, and the occurrence of altercentric biases in one's own perception. We suggest that infants are inherently altercentric in that they preferentially encode events that are co-experienced by another agent. Their own representation of the situation therefore becomes aligned with the agent's perspective, and based on this altercentric representation, they generate correct expectations of how the agent will act. Further, we argue that once infants become aware of themselves as subject of their own actions and perceptual experiences by 2 years of age, these experiences are highlighted, with a similar magnitude as the experience of other agents were before self-emergence. Consequently, they develop an egocentric representation that begins to compete with the prior altercentric representation. This explains why 2- and 3-year-olds' success in false belief tasks is fragile and breaks down once the focus is put on the child's perspective, e.g. with a direct test question (He et al., 2012; Rubio-Fernández & Geurts, 2013), or by emphasizing the child's perception of the object (Rubio-Fernández & Geurts, 2013). While we argue that children encode an altercentric as well as an egocentric perspective by around 2 years of age, until the age of 4 years they lack the executive resources and an appropriate structure for these representations to flexibly select the correct perspective in any given context. This explains why it is not before the age of 4 years that children pass the traditional explicit false belief tasks.

Acknowledgment This work was supported by an ERC Consolidator Grant DEVOMIND 726114 to VS and a Marie Słodowska Curie Fellowship 799734 ToMSelfLink to CGW.

References

Amsterdam, B. (1972). Mirror self-image reactions before age two. *Developmental Psychobiology, 5*(4), 297–305.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953–970. https://doi.org/10.1037/a0016923

Astington, J. W., & Baird, J. A. (2005). *Why language matters for theory of mind*. Oxford: Oxford University Press.

Astington, J. W., & Gopnik, A. (1991). Theoretical explanations of children's understanding of the mind. *British Journal of Developmental Psychology, 9*(1), 7–31. https://doi.org/10.1111/j.2044-835X.1991.tb00859.x

Baillargeon, R. (1986). Representing the existence and the location of hidden objects: Object permanence in 6- and 8-month-old infants. *Cognition, 23*, 21–41.

Baillargeon, R., Devos, J., & Graber, M. (1989). Location memory in 8-month-old infants in a non-search AB task: Further evidence. *Cognitive Development, 4*, 345–367.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14*(3), 110–118. https://doi.org/10.1016/j.tics.2009.12.006

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*(1), 37–46. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9775957

Bennett, J. (1978). Some remarks about concepts - Commentary on cognition and consciousness in nonhuman species. *The Behavioral and Brain Sciences, 4*, 555–629.

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition, 112*(2), 337–342. https://doi.org/10.1016/j.cognition.2009.05.006

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language, 28*(5), 606–637. https://doi.org/10.1111/mila.12036

Carlson, S. M., Claxton, L. J., & Moses, L. J. (2015). The relation between executive function and theory of mind is more than skin deep. *Journal of Cognition and Development, 16*(1), 186–197. https://doi.org/10.1080/15248372.2013.824883

Cunningham, S. J., Brebner, J. L., Quinn, F., & Turk, D. J. (2014). The self-reference effect on memory in early childhood. *Child Development, 85*(2), 808–823. https://doi.org/10.1111/cdev.12144

Dennett, D. C. (1978). Beliefs about beliefs. *The Behavioral and Brain Sciences, 1*, 568–570.

Devine, R. T., & Hughes, C. (2014). Relations between false belief understanding and executive function in early childhood: A meta-analysis. *Child Development, 85*(5), 1777–1794. https://doi.org/10.1111/cdev.12237

Diamond, A. (2012). Executive functions. *Annual Review of Psychology, 64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant theory of mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development, 46*, 12. https://doi.org/10.1016/j.cogdev.2018.01.001

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science, 8*(3), 223–241.

Flavell, J. H., Green, F. L., Flavell, E. R., Watson, M. W., & Campione, J. C. (1986). Development of knowledge about the appearance-reality distinction. *Monographs of the Society for Research in Child Development, 51*(1), 1–87.

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development, 59*(1), 26–37. https://doi.org/10.2307/1130386

Grosse Wiesmann, C. (2017). *The emergence of theory of mind - Cognitive and neural basis of false belief understanding in preschool age* (Vol. 193). Leipzig: Max Planck Institute for Human Cognitive and Brain Sciences.

Grosse Wiesmann, C., Friederici, A. D., Disla, D., Steinbeis, N., & Singer, T. (2018). Longitudinal evidence for 4-year-olds' but not 2- and 3-year-olds' false belief-related action anticipation. *Cognitive Development, 46*(February), 58–68. https://doi.org/10.1016/j.cogdev.2017.08.007

Grosse Wiesmann, C., Friederici, A. D., Singer, T., & Steinbeis, N. (2017). Implicit and explicit false belief development in preschool children. *Developmental Science, 20*(5), 1–15. https://doi.org/10.1111/desc.12445

Grosse Wiesmann, C., Friederici, A. D., Steinbeis, N., & Singer, T. (2020). Two systems for thinking about other people's thoughts in the developing brain. *Proceedings of the National Academy of Sciences of the United States of America, 117*, 6928.

Grosse Wiesmann, C., Schreiber, J., Singer, T., Steinbeis, N., & Friederici, A. D. (2017). White matter maturation is associated with the emergence of theory of mind in early childhood. *Nature Communications, 8*(14692), 1–10. https://doi.org/10.1038/ncomms14692

Goupil, L., Romand-Monnier, M., & Kouider, S. (2016). Infants ask for help when they know they don't know. *Proceedings of the National Academy of Sciences, 113*(13), 3492–3496.

Harman, G. (1978). Studying the chimpanzee's theory of mind - Commentary on Cognition and consciousness in nonhuman species. *Behavioral and Brain Sciences, 4*(1978), 555–629.

He, Z., Bolz, M., & Baillargeon, R. (2012). 2.5-Year-olds succeed at a verbal anticipatory-looking false-belief task. *British Journal of Developmental Psychology, 30*(1), 14–29. https://doi.org/10.1111/j.2044-835X.2011.02070.x

Helming, K. A., Strickland, B., & Jacob, P. (2014). Making sense of early false-belief understanding. *Trends in Cognitive Sciences, 18*, 167–170. https://doi.org/10.1016/j.tics.2014.01.005

Heyes, C. M. (2014). False belief in infancy: A fresh look. *Developmental Science, 17*(5), 647–659. https://doi.org/10.1111/desc.12148

Hogrefe, G.-J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development, 57*(3), 567. https://doi.org/10.2307/1130337

Kampis, D., Somogyi, E., Itakura, S., & Király, I. (2013). Do infants bind mental states to agents? *Cognition, 129*(2). https://doi.org/10.1016/j.cognition.2013.07.004

Knudsen, B., & Liszkowski, U. (2012a). 18-Month-olds predict specific action mistakes through attribution of false belief, not ignorance, and intervene accordingly. *Infancy, 17*(6), 672–691. https://doi.org/10.1111/j.1532-7078.2011.00105.x

Knudsen, B., & Liszkowski, U. (2012b). Eighteen- and 24-month-old infants correct others in anticipation of action mistakes. *Developmental Science, 15*(1), 113–122. https://doi.org/10.1111/j.1467-7687.2011.01098.x

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science (New York, N.Y.), 330*(6012), 1830–1834. https://doi.org/10.1126/science.1190792

Kulke, L., & Rakoczy, H. (2018). Implicit theory of mind–An overview of current replications and non-replications. *Data in Brief, 16*, 101–104.

Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is implicit theory of mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science, 29*(6), 888. https://doi.org/10.1177/0956797617747090

Király, I., Oláh, K., Csibra, G., & Kovács, Á. M. (2018). Retrospective attribution of false beliefs in 3-year-old children. *Proceedings of the National Academy of Sciences, 115*(45), 11477–11482.

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child development, 76*(6), 1261–1277.

Kampis, D., & Kovacs, A. (submitted). Seeing the world from others' perspective: 14-month-olds show altercentric modulation effects by others' beliefs.

Leslie, A. M. (2005). Developmental parallels in understanding minds and bodies. *Trends in Cognitive Sciences, 9*(10), 459–462. https://doi.org/10.1016/j.tics.2005.08.006

Lewis, M., & Ramsay, D. (2004). Development of self-recognition, personal pronoun use, and pretend play during the 2nd year. *Child Development, 75*(6), 1821–1831.

Moll, H., Kane, S., & Mcgowan, L. (2016). Three-year-olds express suspense when an agent approaches a scene with a false belief. *Developmental Science, 19*(2), 208. https://doi.org/10.1111/desc.12310

Moll, H., Khalulyan, A., & Moffett, L. (2017). 2.5-Year-olds express suspense when others approach reality with false expectations. *Child Development, 88*(1), 144. https://doi.org/10.1111/cdev.12581

Musholt, K. (2012). Self-consciousness and intersubjectivity. *Grazer Philosophische Studien, 84*(84), 75–101.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science (New York, N.Y.), 308*(5719), 255–258. https://doi.org/10.1126/science.1107621

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: The MIT Press.

Perner, J. (2011). Theory of mind and levels of understanding. In *Budapest CEU Conference of Cognitive Development*.

Perner, J. (2016). *Referential and cooperative bias: In Defense of an implicit theory of mind*. Retrieved from http://philosophyofbrains.com/2016/10/17/symposium-on-helming-strickland-and-jacob-solving-the-puzzle-about-early-belief-ascription.aspx

Perner, J., & Roessler, J. (2012). From infants' to children's appreciation of belief. *Trends in Cognitive Sciences, 16*(10), 519–525. https://doi.org/10.1016/j.tics.2012.08.004

Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science, 308*, 214–216.

Phillips, J, & Norby, A. (2019) Factive theory of mind. *Mind & Language.* 1– 24. https://doi.org/10.1111/mila.12267

Phillips, J., Ong, D. C., Surtees, A. D. R., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A second look at automatic theory of mind: Reconsidering Kovacs, Teglas, and Endress (2010). *Psychological Science, 26*, 1353–1367. https://doi.org/10.1177/0956797614558717

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development, 46*, 40. https://doi.org/10.1016/j.cogdev.2017.10.004

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(04), 515–526.

Pylyshyn, Z. W. (1978). Cognition and consciousness in honhuman species – Comment. *Behavioral and Brain Sciences, 1*(4), 592–593.

Rochat, P. (2010). The innate sense of the body develops to become a public affair by 2–3 years. *Neuropsychologia, 48*, 738–745. https://doi.org/10.1016/j.neuropsychologia.2009.11.021

Rochat, P., & Striano, T. (1999). Emerging self-exploration by 2-month-old infants. *Developmental Science, 2*, 206–218.

Ross, J., Hutchison, J., & Cunningham, S. J. (2020). The me in memory: The role of the self in autobiographical memory development. *Child Development, 91*, e299–e314. https://doi.org/10.1111/cdev.13211

Rovee-collier, C. K. (1978). Topographical response differentiation and reversal in 3-month-old infants*. *Infant Behavior and Development, 1*, 323–333.

Rubio-Fernández, P., & Geurts, B. (2013). How to pass the false-belief task before your fourth birthday. *Psychological Science, 24*(1), 27–33. https://doi.org/10.1177/0956797612447819

Rubio-Fernández, P., & Geurts, B. (2016). Don't mention the marble! The role of attentional processes in false-belief tasks. *Review of Philosophy and Psychology, 7*(4), 835–850. https://doi.org/10.1007/s13164-015-0290-z

Ruffman, T. (2014). To belief or not belief: Children's theory of mind. *Developmental Review, 34*(3), 265–293. https://doi.org/10.1016/j.dr.2014.04.001

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Scott, S. E. B. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology. Human Perception and Performance, 36*, 1255.

Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences, 21*(4), 237–249. https://doi.org/10.1016/j.tics.2017.01.012

Senju, A., Southgate, V., Snape, C., Leonard, M., & Csibra, G. (2011). Do 18-month-olds really attribute mental states to others? A critical test. *Psychological Science, 22*(7), 878–880. https://doi.org/10.1177/0956797611411584

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in Asperger syndrome. *Science (New York, N.Y.), 325*(5942), 883–885. https://doi.org/10.1126/science.1176170

Setoh, P., Scott, R. M., & Baillargeon, R. (2016). Two-and-a-half-year-olds succeed at a traditional false-belief task with reduced processing demands. *Proceedings of the National Academy of Sciences, 113*, 13360. https://doi.org/10.1073/pnas.1609203113

Sommerville, J. A., Bernstein, D. M., & Meltzoff, A. N. (2013). Measuring beliefs in centimeters: Private knowledge biases preschoolers' and adults' representation of others' beliefs. *Child Development, 84*(6), 1846–1854. https://doi.org/10.1111/cdev.12110

Southgate, V. (2013). Early manifestations of mindreading. In *Understanding other minds: Perspectives from developmental social neuroscience* (pp. 3–18). Oxford: Oxford University Press.

Southgate, V. (2020). Are infants altercentric? The other and the self in early social cognition. *Psychological review, 127*(4), 505.

Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science, 13*(6), 907–912. https://doi.org/10.1111/j.1467-7687.2009.00946.x

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18*(7), 587–592. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20423546

Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *British Journal of Developmental Psychology, 30*(1), 30–44. https://doi.org/10.1111/j.2044-835X.2011.02046.x

Surtees, A., Apperly, I., & Samson, D. (2016). I've got your number: Spontaneous perspective-taking in an interactive task. *Cognition, 150*, 43. https://doi.org/10.1016/j.cognition.2016.01.014

Symons, C. S., Johnson, B. T., Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin, 121*, 371.

Van Der Wel, R. P. R. D., Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition, 130*(1), 128. https://doi.org/10.1016/j.cognition.2013.10.004

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655–684. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11405571

Wilcox, T., Nadel, L., & Rosser, R. (1996). Location memory in healthy preterm and full-term infants. *Infant Behavior and Development, 19*, 309–323.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103–128.

Towards the Integration of Social Cognition and Social Motivation in Autism Spectrum Disorder

Julia Parish-Morris, Robert T. Schultz, and John D. Herrington

Introduction

The neurocognitive mechanisms supporting social behavior—social cognition—have been studied extensively in individuals with autism spectrum disorder (ASD), and deficits in this domain have been argued to underlie the constellation of challenges observed in ASD (Baron-Cohen, 1997; U. Frith, 1989). Nevertheless, the field continues to grapple with widespread individual differences in social cognitive abilities that are not entirely reducible to ASD severity alone. The field also continues to search for theories that account for aspects of the ASD clinical profile that are not adequately explained by social cognitive deficits (for example, repetitive behaviors, sensory differences, and restricted interests). Furthermore, most cognitive neuroscience research in ASD has taken a cross-sectional perspective that assumes a direct or near-direct correspondence between social cognition impairments and

J. Parish-Morris
Department of Psychiatry, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA

Center for Autism Research, Children's Hospital of Philadelphia, Philadelphia, PA, USA

R. T. Schultz
Center for Autism Research, Children's Hospital of Philadelphia, Philadelphia, PA, USA

Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA

J. D. Herrington (✉)
Department of Psychiatry, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA

Center for Autism Research, Children's Hospital of Philadelphia, Philadelphia, PA, USA

Child Psychiatry and Behavioral Science, Roberts Center for Pediatric Research, The Children's Hospital of Philadelphia, Philadelphia, PA, USA
e-mail: herringtonj@email.chop.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_4

neural deficits, without considering the role of development and experience in shaping brain architecture and function. In this chapter, we integrate a review of social cognition in ASD with recently emerging theories of *social motivation* in ASD—and illustrate how this integration promises to enhance our understanding of the etiology and diverse clinical manifestations of ASD.

This chapter begins by briefly reviewing what is meant by "social cognition" in the context of ASD research. Although some researchers define social cognition narrowly, with reference to a fairly circumscribed set of behaviors (e.g., explicit reasoning about people and social situations—as in theory of mind; Premack & Woodruff, 1978), others espouse broader definitions that include detecting, interpreting, and using nonverbal cues like eye gaze, facial expressions, and gesture in social contexts (C. D. Frith, 2008). Despite significant research focusing on specific social cognitive mechanisms in ASD (like face processing or biological motion perception), there is growing recognition that the development of these mechanisms may be strongly coupled with motivational tendencies that are distinct from social cognition as typically construed—i.e., social motivation (Chevallier, Kohls, Troiani, Brodkin, & Schultz, 2012). In this chapter, we first provide a brief review of seminal research studies focused on social cognition, with special attention to theory of mind, face information processing, and biological motion perception. We discuss how these social cognitive impairments in autism might be explained by the social motivation theory of autism, followed by a review of neuroimaging studies of social cognition and social motivation, emphasizing how these generally separate literatures can be theoretically integrated. Finally, we discuss how nonsocial motivations might also explain the pattern of social cognitive deficits observed in ASD, and conclude with recommendations for how to incorporate autistic motivation/reward processing into clinical research and practice.

Social Cognition

Social cognition has received significant experimental attention. As early as the 1970s, researchers began using neuropsychological paradigms to explore patterns of social cognitive strengths and deficits in children with ASD. In 1989, Uta Frith wrote a seminal book that described theory of mind or "mentalizing" as a fundamental impairment in autistic children's ability to attribute mental states (e.g., beliefs, desires, intentions), to themselves and others—and to use those attributions to predict and explain behavior (U. Frith, 1989). A few years later, Simon Baron-Cohen published a paper titled "Mindblindness," which elaborated on this theory by arguing that autism can be explained by specific deficits in the ability to consider other people's thoughts and experiences as separate from one's own (Baron-Cohen, 1997). Social skills deficits typical of ASD (e.g., difficulty understanding sarcasm, failure to adhere to social norms) have typically been attributed to fundamental problems with social cognition.

Theory of Mind

Over the years, a variety of experimental paradigms have been devised to test children's explicit and implicit social cognition, with a focus on Theory of Mind (ToM). The most famous is the Sally & Anne task (Wimmer & Perner, 1983), wherein children observe a sequence of acted events (generally performed by two puppets) that lead to a false belief in one character (Baron-Cohen, Leslie, & Frith, 1985). For example, Puppet A puts an object in one of two boxes and leaves the scene. Puppet B comes into the scene and moves the object from one box to the other. Puppet A returns to get the object and children are asked to point to the box where Puppet A will look. Success on this task requires children to realize that Puppet A is unaware that Puppet B has changed the location of the object, and that Puppet A will therefore not have correct knowledge (i.e., they will have a *false belief* about the location of the object). When asked where Puppet A will look for the object, children must point to the original box—thus overriding their own knowledge of the true location of the object. This ability to hold true information in one's mind while attributing a false belief to someone else is a challenging task that typically developing (TD) children cannot reliably implement until they are 4 years or older (Mitchell, 1997). Children with ASD often solve this task significantly later than TD children (Baron-Cohen et al., 1985), and performance on ToM tasks has been linked to a variety of other deficit areas, including language (Andrés-Roqueta & Katsos, 2017) and social skill (Peterson, Slaughter, Moore, & Wellman, 2016). Although accounts vary, most of them share in common the premise that ToM is at least somewhat modular in function—i.e., it is implemented by dedicated cognitive modules and corresponding brain circuits (C. D. Frith & Frith, 2006).

Theory of mind deficits have been invoked to explain a variety of challenges within ASD, many of which are not as simple as reasoning about other people's thoughts and feelings. For example, ToM is necessary to detect and respond to other people's intentions and goals, which in turn forms an important foundation for other higher-order social cognitive behaviors, like understanding and making jokes, lying, and comprehending irony and sarcasm (Channon, Pellijeff, & Rule, 2005). When viewing ASD through a primarily social lens, many symptoms can be parsimoniously explained by the presence of a specific social cognitive deficit, as in the case of individuals who are highly intelligent but still have difficulty interacting socially.

Face Perception

The social cognitive view of ASD suggests that a variety of social deficits can be traced to underlying cognitive impairments. For example, the ability to recognize and interpret facial information (including expressions and identity information) is an important part of social life, and has shown to be diminished in ASD (Loukusa, Mäkinen, Kuusikko-Gauffin, Ebeling, & Moilanen, 2014; Shanok, Jones, & Lucas, 2019; Wolf et al., 2008). Intact facial recognition processing is central to successful ToM, insofar as faces provide critical information about the thoughts and feelings of others.

Biological Motion Perception

Another key component of social cognitive models of ASD is related to the perception of biological motion. The visual system contains modules that are specifically tuned to the perception of human movement (for review, see Allison, Puce, & McCarthy, 2000). Deficits in biological motion perception are viewed as an explanation for the difficulties many autistic individuals have in interpreting nonverbal social communicative cues (namely gestures).

In evaluating their impact as social cognitive models of ASD, it is important to point out that facial and biological motion information processes overlap theoretically and psychophysically (Thompson, Hardee, Panayiotou, Crewther, & Puce, 2007). In particular, both require the analysis of configural relationships between features (as opposed to individual features alone)—which individuals with ASD often find challenging (this is a key tenet of the central coherence theory of ASD; Shah & Frith, 1993). In fact, as purely visual input, the highly configural nature of biological entities (including animals; see Mather & West, 1993) is arguably the most meaningful difference between biological and non-biological visual perception.

ToM, face processing, and biological motion perception have often been studied in isolation, but they tend to share a fundamental premise that the social challenges experienced by individuals with ASD can be understood in terms of modular deficits in cognitive information processing. An alternative proposal—the social motivation hypothesis of ASD—suggests that observed differences in social cognition may not relate to the integrity of cognitive functions per se (like face processing), but instead, relate to the manner in which these processes are (or are not) implemented or prioritized across development.

Social Motivation

In 2012, Chevallier and colleagues argued that diminished social motivation (Dawson et al., 2002; Dawson, Webb, & McPartland, 2005) could provide a more comprehensive yet parsimonious account of ASD than social cognitive deficits. Described as "a set of psychological dispositions and biological mechanisms biasing the individual to preferentially orient to the social world (social orienting), to seek and take pleasure in social interactions (social reward), and to work to foster and maintain social bonds (social maintaining)," social motivation was posited to be "an evolutionary adaptation geared to enhance the individual's fitness in collaborative environments." To support their argument, the authors drew on diverse evidence from the fields of social psychology, behavioral economics, social neuroscience, and evolutionary biology. One of the most significant arguments made by Chevallier and colleagues was that deficits in social cognition writ large—including ToM, face processing, and biological motion impairments—are the *consequence*, or downstream effect, of diminished social motivation (Chevallier et al., 2012).

The impact of this argument is illustrated by our evolving view of diminished preference for face stimuli in ASD. Diminished attention to faces in ASD has been cited as strong evidence for the social motivation hypothesis, because a powerful draw to human faces and face-like schemata is among the earliest behaviors observed in human infants. Early face biases make sense from an evolutionary perspective; given that infants must rely on others to care for them, it is logical that they are born with an adaptive bias to attend to stimuli most likely to provide them with food and warmth. Infants indeed show early preferences for faces and face-like stimuli (Goren, Sarty, & Wu, 1975). Using newer technologies, some have argued that a bias towards attending to face-like stimuli can be measured in fetuses during the third trimester of pregnancy (Reid et al., 2017), indicating that experience with human faces is not necessary to trigger face biases. Thus, infants may be born with innate biases that "prepare" them to seek stimuli associated with survival.

Autism is not typically diagnosed until after age 4, which makes it challenging to study the emergence of face biases in this population. However, the first signs of ASD can be detected by 12 months in high-risk samples (IBIS network et al., 2015), and trained experts can diagnose ASD with reasonable accuracy around age 2 (Ben Itzchak & Zachor, 2009; Ozonoff et al., 2015). To circumvent the problem of late diagnosis, researchers have leveraged family samples. In these "high-risk" infant sibling studies, scientists recruit families that already have one child with ASD and are pregnant with another child. ASD is estimated to occur in approximately 1.5% of the population (Baio, 2018), but the likelihood of having a second child with ASD is much higher (Georgiades et al., 2013). Infant sibling studies are therefore enriched for infants with an eventual ASD diagnosis, and provide researchers with the opportunity to study how the condition unfolds from birth through diagnosis.

Studies of face processing in infants who later develop ASD suggest very early disruptions in innate face biases. For example, high-risk infants who are shown videos with social stimuli on one side of the screen (children smiling and playing) or colorful fractals on the other side of the screen will attend more to the fractals than low-risk infants (Pierce et al., 2016). As infants become toddlers, they remain less interested in social stimuli (and more drawn to nonsocial stimuli) than non-ASD peers (Chawarska, Macari, & Shic, 2012), and thus acquire less experience with faces. This relative lack of experience with faces has been argued to cause deficits in higher level face processing tasks, such as recognizing and interpreting facial emotions.

Nuances of the Social Motivation Theory

The social motivation theory of ASD has several advantages over traditional social cognitive models of ASD. In particular, it helps to explain why symptoms of ASD manifest long before the development of social cognitive milestones (i.e., in infancy, long before the maturation of ToM). It also may do a better job of explaining individual differences in ASD symptom profiles, insofar as social motivation deficits may vary in their impact on specific cognitive functions; motivation may mediate the development of social cognition, but not perfectly, and not necessarily in the same way across all aspects of social cognition. However, the social motivation theory does have significant limitations that warrant consideration and future study.

First and foremost, social motivation is not universally impaired in ASD, which is consistent with the notion that autism is composed of several distinct conditions with unique causes and phenotypes. Recent studies have shown that girls with ASD, in particular, may be characterized by higher social motivation than otherwise comparable boys (Sedgewick, Hill, Yates, Pickering, & Pellicano, 2016). Repeated failures to establish social bonds, despite motivation to make friends, may contribute to experiences of frustration, depression, and anxiety. The fundamental problem experienced by these individuals is not lack of social motivation, but rather, a difficulty in translating social motivation into social skills. For people with this social profile, social motivation does not predict social success—which can lead to significant distress.

It is also less clear how the social motivation hypothesis explains the reward value many individuals with ASD attribute to circumscribed special interests and repetitive behaviors. One possible pathway from diminished social motivation to the special interests/repetitive behaviors observed in ASD is via nonsocial specialization. On this view, reductions in biases towards attending to social stimuli might lead—almost by default—to unusual amounts of attention allocated to nonsocial stimuli (e.g., toys, fans), thus inadvertently "specializing" the brain for nonsocial purposes rather than for social purposes. This possibility, although plausible, may not be the most parsimonious explanation for patterns of nonsocial attention and specialization observed in ASD. Rather, an emerging body of research suggests that atypical nonsocial motivations and reward processes could more concisely explain aspects of the ASD phenotype. For example, biases towards attending to perceptually salient objects or actions (e.g., fractals; Pierce et al., 2016) could lead to reduced social experience, with similar downstream effects as would be predicted by the social motivation hypothesis. Thus, it is possible that the ASD phenotype stems from increased motivation to attend to nonsocial stimuli, rather than decreased motivation to attend to social stimuli (Gale, Eikeseth, & Klintwall, 2019).

There is growing recognition that differences in reward responsiveness in ASD may extend well beyond social perception and behavior, and may not conform strictly to a deficit model. In particular, it is difficult to reconcile the diminished social reward typically observed in ASD with the heightened reward responsiveness to things like circumscribed special interests. A focus on social rewards alone may prove to be an oversimplification of the motivational differences that characterize individuals with ASD. Here we encounter fundamental limitations of our theoretical understanding of what constitute reward; neurobiological models of reward motivation (discussed below) have arguably not gone far enough in differentiating between different types of reward, and the extent to which they do or do not map onto dissociable mechanisms (for discussion, see Clements et al., 2018; Kohls, Schulte-Rüther, et al., 2012). The future of the social motivation hypothesis is likely to turn on the extent to which it adequately captures the complete profile of motivational dispositions in ASD.

The Cognitive Neuroscience of Social Motivation and Social Information Processing

Neural systems associated with the motivation and reward mechanisms discussed above have been the subject of decades of human and preclinical research. The same is true of the neural systems supporting social cognition. There are many reviews of both literatures, including reviews that detail how these systems function differently in autistic individuals (for some highlights see Adolphs, 2009; Berridge & Kringelbach, 2015; Ikemoto, 2010; Kohls, Chevallier, Troiani, & Schultz, 2012; Schultz, 2005). However, these literatures are seldom integrated—which is unfortunate, as their integration reveals some of the fundamental principles that shape the development of social behavior in typical development and in ASD.

The Reward System

The neural system associated with reward is primarily subcortical, following what is known as the mesolimbic pathway (Lammel, Lim, & Malenka, 2014; O'Connell & Hofmann, 2011). This pathway connects midbrain structures (ventral tegmental area, or VTA) to the thalamus and forebrain structures within the basal ganglia. The striatum is a major subdivision of the basal ganglia, which is itself divided into subregions that play somewhat distinct roles in reward (for review see Lenz & Lobo, 2013). Key reward structures within dorsal branch of the striatum include the caudate nucleus and putamen; the ventral striatum includes the nucleus accumbens. Although these areas represent the core of the reward system, this system extends to other subcortical and cortical structures, including amygdala and portions of prefrontal cortex (particularly orbitofrontal and anterior cingulate cortices; for review, see Kohls, Chevallier, et al., 2012).

The relay mechanisms between structures of the reward system are also well known. The mesolimbic system consists of a dopaminergic pathway whereby the VTA sends inputs to the nucleus accumbens (among other structures; Ikemoto, 2010). The dopaminergic mechanisms of this system interact with several other mechanisms and neurotransmitters, including opioids, GABAergic, glutaminergic, and cholinergic neurons and interneurons (R. C. Pierce & Kumaresan, 2006).

While different types of reward (social reward, money, etc.) are thought to share this common pathway (Clements et al., 2018), there is likely some diversity among the structures in terms of the specific components of reward they implement. Although there are several theoretical accounts of reward, a major distinction is typically made between the anticipation of reward ("wanting"), and the consumption of reward ("liking"; Berridge, Robinson, & Aldridge, 2009). These constructs have significant anatomical, conceptual, and experiential overlap (i.e., the experience of liking reinforces wanting, and vice versa). But they appear to be at least partially distinguishable and have different implications for neurodevelopmental conditions like ASD. To date, the preponderance of human neuroimaging evidence suggests difference among autistic individuals in the anticipatory components of social contact ("wanting"), and the structures associated with these components (in particular, ventral striatum). But the evidence for differences in social reward *consumption* ("liking") in ASD are presently less clear (for reviews, see Clements et al., 2018; Kohls, Chevallier, et al., 2012). This is an important area of research, as many autistic individuals report negative emotional responses to diminished social contact (Gotham, Bishop, Brunwasser, & Lord, 2014; Hedley, Uljarević, Wilmot, Richdale, & Dissanayake, 2018), despite differences in social approach motivation. This suggests intact "liking" despite diminished "wanting." Thus, some components of social reward, but not others, may be affected in ASD; which components are affected may itself represent an important individual difference variable (Chevallier et al., 2012; Kohls, Chevallier, et al., 2012) (Fig. 1).

The reward system extends beyond the striatum to other brain areas that are known to play a role in social cognition, and also in ASD more broadly. In particular,

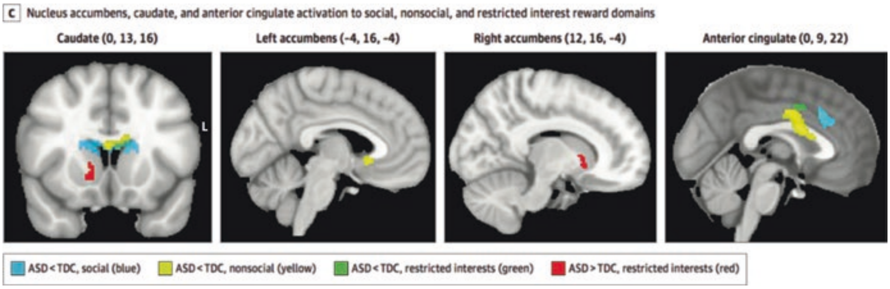

C
Nucleus accumbens, caudate, and anterior cingulate activation to social, nonsocial, and restricted interest reward domains
Caudate (0, 13, 16)
Left accumbens (-4, 16, -4)
Right accumbens (12, 16, -4)
Anterior cingulate (0, 9, 22)
L
🟦 ASD < TDC, social (blue)
🟨 ASD < TDC, nonsocial (yellow)
🟩 ASD < TDC, restricted interests (green)
🟥 ASD > TDC, restricted interests (red)

Fig. 1 Differences in reward system activity in ASD. *Note.* Differences in reward system activity for social and nonsocial rewards in ASD. In the caudate, individuals with ASD showed hypoactivation to social stimuli (blue), nonsocial stimuli (yellow), and hyperactivation to restricted interest stimuli (red) compared with controls. In the nucleus accumbens, individuals with ASD showed hypoactivation in the right hemisphere to nonsocial stimuli (yellow) and hyperactivation in the left to restricted interests (red). No significant cluster involving the accumbens was observed in the social meta-analytic results. In the anterior cingulate cortex, individuals with ASD showed hypoactivation to social stimuli, nonsocial stimuli, and restricted interest stimuli, compared with controls. RE indicates random effects. (Taken from Clements et al., 2018)

amygdala plays a critical role in reward-based learning, alongside its role in the perception and experience of affect (for review see Baxter & Murray, 2002). The reward system also extends to anterior cingulate cortex (ACC), which plays a role in response selection and decision-making in the context of reward (Bush et al., 2002). Although traditional ToM regions are often localized adjacent to ACC, there is some evidence that they include portions of ACC as well (for review, see Apps, Rushworth, & Chang, 2016). Furthermore, there is some evidence that projections from the mesolimbic areas of the reward system extend beyond ACC to adjacent ToM areas (Supekar et al., 2018).

Social Cognition and the "Social Brain"

With the noteworthy exceptions of amygdala and ventral prefrontal cortex, most of the areas traditionally associated with social information processing are located outside of the reward system. These areas include temporal visual information processing structures—in particular, fusiform gyrus (FG, associated with facial information processing; Schultz et al., 2003) and superior temporal sulcus (STS, associated with biological motion perception; Thompson, Clarke, Stewart, & Puce, 2005; note that, because these literatures are vast, we limit ourselves to one general reference per structure). They also include structures associated with ToM and action understanding, including the temporo-parietal junction (Saxe, 2006) and portions of medial prefrontal cortex (U. Frith & Frith, 2003). Two decades of cognitive neuroscience research have shown that these areas function differently in ASD, and that these differences are associated to varying degrees with social cognitive deficits that are common in ASD.

The Integration of Reward and Social Cognition Systems

It is worth noting that the social motivation hypothesis of ASD involves a neural system that is largely distinct from the "social brain" that has been the primary focus of two decades of ASD cognitive neuroscience research. This begs the question of precisely how deficits in social reward mechanisms relate to social cognitive mechanisms. Although empirical data on this relationship are thin, three general accounts seem plausible (none being mutually exclusive). The first is that both cortical (social brain) and subcortical (reward) structures are affected in ASD by a shared neurodevelopmental deficit. This is consistent with models of ASD brain development that emphasize global mechanisms (for example, gray/white matter overgrowth theories, and theories regarding neuronal pathfinding; Ke et al., 2009; Sacco, Gabriele, & Persico, 2015; Zhang et al., 2018). A second possible relationship is that reward and social cognition brain structures influence one another directly, via neural connectivity (for an overview on prefrontal modulation of reward mechanisms, see Banich & Floresco, 2019). We now know that "top-down" and "bottom-up" influences between cortical and subcortical structures are commonplace—for example, subcortical structures involved in emotion and reward tune both frontal (ToM) and visual cortical structures (face and biological motion processing, Gee et al., 2013; Herrington, Nymberg, Faja, Price, & Schultz, 2012; Herrington, Taylor, Grupe, Curby, & Schultz, 2011; Kim et al., 2011). It is therefore possible that deficits in inputs from reward structures lead to diminished function of structures involved in social cognition (while the effect may also operate in the opposite direction, the development of basic reward mechanisms is likely to precede many higher-order social cognitive mechanisms).

A third possible relationship between reward and social cognition steps outside of traditional cognitive neuroscience research in a fundamental and important way. In short, the relationship between neural systems involved in social motivation and cognition is likely to be heavily mediated by *experience*. Perhaps the most significant consequence of diminished social motivation is a decrease in opportunities to provide inputs to social cognitive systems that need those inputs to mature and specialize.

This phenomenon can be illustrated most clearly in the context of facial information processing. While face processing relies a distributed network of brain areas, the most prominent among these is the Fusiform Face Area (FFA) within fusiform gyrus (FG). The face-specific tuning of this area follows a developmental progression across childhood (Golarai, Ghahremani, Grill-Spector, & Gabrieli, 2005; Gomez et al., 2017; Zhu, Bhatt, & Joseph, 2016), such that children become better at face processing over time. While the face-specific nature of this tuning is sometimes regarded as partially innate (as discussed above), an important body of literature suggests that FG activity relates to expert configural processing of *any* visual stimulus (examples from this literature include the expert processing of birds and cars; Bilalić, 2016; Gauthier, Skudlarski, Gore, & Anderson, 2000; McGugin, Gatenby, Gore, & Gauthier, 2012). In other words, FFA activity tracks improvement in visual recognition acquired from experience.

For neurotypical individuals, visual experience begins at or near birth, and is focused on faces more than any other type of stimulus. Infants show clear preferences for facial information (Dalrymple et al., 2018; Farroni et al., 2005; Frank, Vul, & Johnson, 2009), even at stages of neurodevelopment where visual information as a whole is impoverished (acuity, depth perception, and color vision mature significantly in the first few months of life; Adams, 1987; Bronson, 1990; Courage & Adams, 1990; Kavsek, Granrud, & Yonas, 2009). Although functional imaging data on newborns is scarce and challenging to acquire, it seems likely that newborn preferences for facial information relate to components of the neural reward system.

On the other hand, newborns who later carry an ASD diagnosis show a diminished preference for facial information (Chawarska, Macari, & Shic, 2013; Elsabbagh et al., 2013; Shic, Macari, & Chawarska, 2014) that, for many individuals on the spectrum, persists into adulthood (for review see Chita-Tegmark, 2016). The literature on FFA and visual expertise directly predicts that individuals who spend less time and effort processing faces will show decreased face-specific activity in this area. In other words, decreased social motivation may lead to impaired social cognitive processing not solely due to diminished neural functioning or connectivity, but rather, because brain development was shaped—interactively and in a cascading manner—by atypical perceptual experiences and behavior.

This leads to the intriguing possibility that some social cognitive deficits in ASD are not directly related to deficits in the structure or function of underlying brain areas. An illustration of this phenomenon comes a case study from Grelotti et al. (2005), who showed increased FFA activation for visual information that was highly relevant (and rewarding—cartoon characters), but not for faces. In other words, autistic individuals may engage "social" cognitive mechanisms to process information that neurotypicals do not. The social motivation hypothesis thereby provides a significant, and arguably less reductionistic, reframing of what constitutes a social cognitive deficit in ASD; instead of focusing on the diminished functioning of mechanisms, it focuses on differential sensitivity of these mechanisms for specific perceptual experiences.

Conclusion

Studies of the etiology of ASD have undergone several theoretical iterations in recent decades, including cogent arguments that single explanations of ASD are destined to be fruitless (Happé, Ronald, & Plomin, 2006). When surveyed broadly, the different theories of social cognitive and motivational differences in ASD seem to integrate in ways that are not readily apparent at any given moment in time. We contend that the emergence of motivational theories of ASD (the social motivation hypothesis) offers a vastly expanded perspective on social cognition in ASD—one that moves away from core deficit models of social behavior, and towards a dynamic model of how experience mediates the development and implementation of social life.

There remain several challenges inherent in motivational and experiential accounts of social cognitive development. While explanations of ASD that hinge on deficits in social motivation are supported by emerging brain-behavior literature, both social cognitive and motivational accounts are challenged to explain the repetitive behaviors and restricted patterns of interests observed in autism. Atypical motivation and reward theories that allow for an active nonsocial component might characterize autism more fully, and an integrated social/nonsocial approach is likely to bear the most fruit. Future research in this area is sorely needed, particularly longitudinal brain-behavior studies that integrate multiple methods in large and diverse cohorts, and intervention research designed to elucidate mechanisms of change.

While much of the cognitive neuroscience literature reviewed above rests on a large body of empirical data, the big-picture integration of social motivation and cognitive accounts of ASD remains speculative. Formal tests of this integration will require a fundamentally different approach to research than is typically afforded by cross-sectional studies of brain and behavior. The effects of social motivation on the development of cognitive systems is likely to play out over years, in ways that are highly interactive and non-linear. Longitudinal brain and behavior studies are required, and even these will face challenges in inferring causality (i.e., whether diminished social motivation results in diminished social contact, which *causes* diminished social cognition). The relationship between social motivation and cognition may also vary significantly depending on the cognitive mechanism under investigation. While the narrative surrounding the expert specialization of FFA may have significant empirical support, it is less clear that this type of support exists for, say, the specialization of STS for biological motion perception or medial prefrontal cortex for ToM. Ultimately, the identification of data linking social motivation to cognitive mechanisms would go a long way towards validating motivational accounts of ASD.

Treatment research also has tremendous potential in elucidating how social motivation and cognition interact. A major implication of motivation-driven models of social cognition is that processes such as ToM, face processing, and biological motion perception can be shaped by changes in behavior and experience—precisely the changes sought in treatment. But motivation is itself a driving force behind behavioral change, and the absence of such motivation is very difficult to compensate for in treatment (this is arguably why intervention modalities such as applied behavior analysis are so time- and resource-intensive; children are otherwise disinclined to participate). An increased emphasis on cognitive mechanisms as treatment outcomes (alongside more traditional targets like challenging behaviors) will be critical in testing theories on the primacy of social motivation in shaping social cognition.

References

Adams, R. J. (1987). An evaluation of color preference in early infancy. *Infant Behavior and Development, 10*(2), 143–150. https://doi.org/10.1016/0163-6383(87)90029-4

Adolphs, R. (2009). The social brain: Neural basis of social knowledge. *Annual Review of Psychology, 60*(1), 693–716. https://doi.org/10.1146/annurev.psych.60.110707.163514

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. *Trends in Cognitive Sciences, 4*(7), 267–278.

Andrés-Roqueta, C., & Katsos, N. (2017). The contribution of grammar, vocabulary and theory of mind in pragmatic language competence in children with autistic spectrum disorders. *Frontiers in Psychology, 8*, 996. https://doi.org/10.3389/fpsyg.2017.00996

Apps, M. A. J., Rushworth, M. F. S., & Chang, S. W. C. (2016). The anterior cingulate gyrus and social cognition: Tracking the motivation of others. *Neuron, 90*(4), 692–707. https://doi.org/10.1016/j.neuron.2016.04.018

Baio, J. (2018). Prevalence of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveillance Summaries, 67*, 1. https://doi.org/10.15585/mmwr.ss6706a1

Banich, M. T., & Floresco, S. (2019). Reward systems, cognition, and emotion: Introduction to the special issue. *Cognitive, Affective, & Behavioral Neuroscience, 19*, 409. https://doi.org/10.3758/s13415-019-00725-z. s13415-019-00725-z [pii].

Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "Theory of Mind"? *Cognition, 21*, 37–46.

Baxter, M. G., & Murray, E. A. (2002). The amygdala and reward. *Nature Reviews. Neuroscience, 3*(7), 563–573. https://doi.org/10.1038/nrn875

Ben Itzchak, E., & Zachor, D. A. (2009). Change in autism classification with early intervention: Predictors and outcomes. *Research in Autism Spectrum Disorders, 3*(4), 967–976. https://doi.org/10.1016/j.rasd.2009.05.001

Berridge, K. C., & Kringelbach, M. L. (2015). Pleasure systems in the brain. *Neuron, 86*(3), 646–664. https://doi.org/10.1016/j.neuron.2015.02.018

Berridge, K. C., Robinson, T. E., & Aldridge, J. W. (2009). Dissecting components of reward: 'Liking', 'wanting', and learning. *Current Opinion in Pharmacology, 9*(1), 65–73. https://doi.org/10.1016/j.coph.2008.12.014

Bilalić, M. (2016). Revisiting the role of the fusiform face area in expertise. *Journal of Cognitive Neuroscience, 28*(9), 1345–1357. https://doi.org/10.1162/jocn_a_00974

Bronson, G. W. (1990). Changes in infants' visual scanning across the 2- to 14-week age period. *Journal of Experimental Child Psychology, 49*(1), 101–125.

Bush, G., Vogt, B. A., Holmes, J., Dale, A. M., Greve, D., Jenike, M. A., & Rosen, B. R. (2002). Dorsal anterior cingulate cortex: A role in reward-based decision making. *Proceedings of the National Academy of Sciences of the United States of America, 99*(1), 523–528. https://doi.org/10.1073/pnas.012470999

Channon, S., Pellijeff, A., & Rule, A. (2005). Social cognition after head injury: Sarcasm and theory of mind. *Brain and Language, 93*(2), 123–134. https://doi.org/10.1016/j.bandl.2004.09.002

Chawarska, K., Macari, S., & Shic, F. (2012). Context modulates attention to social scenes in toddlers with autism: Context modulates social attention in autism. *Journal of Child Psychology and Psychiatry, 53*(8), 903–913. https://doi.org/10.1111/j.1469-7610.2012.02538.x

Chawarska, K., Macari, S., & Shic, F. (2013). Decreased spontaneous attention to social scenes in 6-month-old infants later diagnosed with autism spectrum disorders. *Biological Psychiatry, 74*(3), 195–203. https://doi.org/10.1016/j.biopsych.2012.11.022

Chevallier, C., Kohls, G., Troiani, V., Brodkin, E. S., & Schultz, R. T. (2012). The social motivation theory of autism. *Trends in Cognitive Sciences, 16*(4), 231–239. https://doi.org/10.1016/j.tics.2012.02.007

Chita-Tegmark, M. (2016). Social attention in ASD: A review and meta-analysis of eye-tracking studies. *Research in Developmental Disabilities, 48*, 79–93. https://doi.org/10.1016/j.ridd.2015.10.011

Clements, C. C., Zoltowski, A. R., Yankowitz, L. D., Yerys, B. E., Schultz, R. T., & Herrington, J. D. (2018). Evaluation of the social motivation hypothesis of autism: A systematic review and meta-analysis. *JAMA Psychiatry, 75*(8), 797–808. https://doi.org/10.1001/jamapsychiatry.2018.1100

Courage, M. L., & Adams, R. J. (1990). Visual acuity assessment from birth to three years using the acuity card procedure: Cross-sectional and longitudinal samples. *Optometry and Vision Science: Official Publication of the American Academy of Optometry, 67*(9), 713–718.

Dalrymple, K. A., Wall, N., Spezio, M., Hazlett, H. C., Piven, J., & Elison, J. T. (2018). Rapid face orienting in infants and school-age children with and without autism: Exploring measurement invariance in eye-tracking. *PLoS One, 13*(8), e0202875. https://doi.org/10.1371/journal.pone.0202875

Dawson, G., Carver, L., Meltzoff, A. N., Panagiotides, H., McPartland, J., & Webb, S. J. (2002). Neural correlates of face and object recognition in young children with autism spectrum disorder, developmental delay, and typical development. *Child Development, 73*(3), 700–717. https://doi.org/10.1111/1467-8624.00433

Dawson, G., Webb, S. J., & McPartland, J. (2005). Understanding the nature of face processing impairment in autism: Insights from behavioral and electrophysiological studies. *Developmental Neuropsychology, 27*(3), 403–424. https://doi.org/10.1207/s15326942dn2703_6

Elsabbagh, M., Gliga, T., Pickles, A., Hudry, K., Charman, T., & Johnson, M. H. (2013). The development of face orienting mechanisms in infants at-risk for autism. *Behavioural Brain Research, 251*, 147–154. https://doi.org/10.1016/j.bbr.2012.07.030

Farroni, T., Johnson, M. H., Menon, E., Zulian, L., Faraguna, D., & Csibra, G. (2005). Newborns' preference for face-relevant stimuli: Effects of contrast polarity. *Proceedings of the National Academy of Sciences, 102*(47), 17245–17250. https://doi.org/10.1073/pnas.0502205102

Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition, 110*(2), 160–170. https://doi.org/10.1016/j.cognition.2008.11.010

Frith, C. D. (2008). Social cognition. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 363*(1499), 2033–2039. https://doi.org/10.1098/rstb.2008.0005

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50*(4), 531–534. https://doi.org/10.1016/j.neuron.2006.05.001

Frith, U. (1989). Autism and "Theory of Mind". In C. Gillberg (Ed.), *Diagnosis and treatment of autism* (pp. 33–52). New York, NY: Springer. https://doi.org/10.1007/978-1-4899-0882-7_4

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 358*(1431), 459–473. https://doi.org/10.1098/rstb.2002.1218

Gale, C. M., Eikeseth, S., & Klintwall, L. (2019). Children with autism show atypical preference for non-social stimuli. *Scientific Reports, 9*, 10355. https://doi.org/10.1038/s41598-019-46705-8

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience, 3*(2), 191–197. https://doi.org/10.1038/72140

Gee, D. G., Humphreys, K. L., Flannery, J., Goff, B., Telzer, E. H., Shapiro, M., … Tottenham, N. (2013). A developmental shift from positive to negative connectivity in human amygdala-prefrontal circuitry. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 33*(10), 4584–4593. https://doi.org/10.1523/JNEUROSCI.3446-12.2013

Georgiades, S., Szatmari, P., Zwaigenbaum, L., Bryson, S., Brian, J., Roberts, W., … Garon, N. (2013). A prospective study of autistic-like traits in unaffected siblings of probands with autism spectrum disorder. *JAMA Psychiatry, 70*(1), 42–48. https://doi.org/10.1001/2013.jamapsychiatry.1

Golarai, G., Ghahremani, D. G., Grill-Spector, K., & Gabrieli, J. D. E. (2005). Evidence for maturation of the fusiform face area (FFA) in 7 to 16 year old children. *Journal of Vision, 5*(8), 634–634. https://doi.org/10.1167/5.8.634

Gomez, J., Barnett, M. A., Natu, V., Mezer, A., Palomero-Gallagher, N., Weiner, K. S., … Grill-Spector, K. (2017). Microstructural proliferation in human cortex is coupled with the development of face processing. *Science (New York, N.Y.), 355*(6320), 68–71. https://doi.org/10.1126/science.aag0311

Goren, C. C., Sarty, M., & Wu, P. Y. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics, 56*(4), 544–549.

Gotham, K., Bishop, S. L., Brunwasser, S., & Lord, C. (2014). Rumination and perceived impairment associated with depressive symptoms in a verbal adolescent-adult ASD sample: Depression in ASD. *Autism Research, 7*(3), 381–391. https://doi.org/10.1002/aur.1377

Grelotti, D. J., Klin, A. J., Gauthier, I., Skudlarski, P., Cohen, D. J., Gore, J. C., … Schultz, R. T. (2005). FMRI activation of the fusiform gyrus and amygdala to cartoon characters but not to faces in a boy with autism. *Neuropsychologia, 43*(3), 373–385.

Happé, F., Ronald, A., & Plomin, R. (2006). Time to give up on a single explanation for autism. *Nature Neuroscience, 9*(10), 1218–1220. nn1770 [pii].

Hedley, D., Uljarević, M., Wilmot, M., Richdale, A., & Dissanayake, C. (2018). Understanding depression and thoughts of self-harm in autism: A potential mechanism involving loneliness. *Research in Autism Spectrum Disorders, 46*, 1–7. https://doi.org/10.1016/j.rasd.2017.11.003

Herrington, J. D., Nymberg, C., Faja, S., Price, E., & Schultz, R. T. (2012). The responsiveness of biological motion processing areas to selective attention towards goals. *NeuroImage, 63*(1), 581–590. https://doi.org/10.1016/j.neuroimage.2012.06.077

Herrington, J. D., Taylor, J. M., Grupe, D. W., Curby, K. M., & Schultz, R. T. (2011). Bidirectional communication between amygdala and fusiform gyrus during facial recognition. *NeuroImage, 56*(4), 2348–2355. https://doi.org/10.1016/j.neuroimage.2011.03.072

IBIS Network, Estes, A., Zwaigenbaum, L., Gu, H., St. John, T., Paterson, S., … Piven, J. (2015). Behavioral, cognitive, and adaptive development in infants with autism spectrum disorder in the first 2 years of life. *Journal of Neurodevelopmental Disorders, 7*(1). https://doi.org/10.1186/s11689-015-9117-6

Ikemoto, S. (2010). Brain reward circuitry beyond the mesolimbic dopamine system: A neurobiological theory. *Neuroscience and Biobehavioral Reviews, 35*(2), 129–150. https://doi.org/10.1016/j.neubiorev.2010.02.001

Kavsek, M., Granrud, C. E., & Yonas, A. (2009). Infants' responsiveness to pictorial depth cues in preferential-reaching studies: A meta-analysis. *Infant Behavior & Development, 32*(3), 245–253. https://doi.org/10.1016/j.infbeh.2009.02.001

Ke, X., Tang, T., Hong, S., Hang, Y., Zou, B., Li, H., … Liu, Y. (2009). White matter impairments in autism, evidence from voxel-based morphometry and diffusion tensor imaging. *Brain Research, 1265*, 171–177.

Kim, M. J., Loucks, R. A., Palmer, A. L., Brown, A. C., Solomon, K. M., Marchante, A. N., & Whalen, P. J. (2011). The structural and functional connectivity of the amygdala: From normal emotion to pathological anxiety. *Behavioural Brain Research, 223*(2), 403–410. https://doi.org/10.1016/j.bbr.2011.04.025

Kohls, G., Chevallier, C., Troiani, V., & Schultz, R. T. (2012). Social "wanting" dysfunction in autism: Neurobiological underpinnings and treatment implications. *Journal of Neurodevelopmental Disorders, 4*(1), 10. https://doi.org/10.1186/1866-1955-4-10

Kohls, G., Schulte-Rüther, M., Nehrkorn, B., Müller, K., Fink, G. R., Kamp-Becker, I., … Konrad, K. (2012). Reward system dysfunction in autism spectrum disorders. *Social Cognitive and Affective Neuroscience, 8*(5), 565–572. https://doi.org/10.1093/scan/nss033

Lammel, S., Lim, B. K., & Malenka, R. C. (2014). Reward and aversion in a heterogeneous midbrain dopamine system. *Neuropharmacology, 76*, 351–359. https://doi.org/10.1016/j.neuropharm.2013.03.019

Lenz, J. D., & Lobo, M. K. (2013). Optogenetic insights into striatal function and behavior. *Behavioural Brain Research, 255*, 44–54. https://doi.org/10.1016/j.bbr.2013.04.018

Loukusa, S., Mäkinen, L., Kuusikko-Gauffin, S., Ebeling, H., & Moilanen, I. (2014). Theory of mind and emotion recognition skills in children with specific language impairment, autism spectrum disorder and typical development: Group differences and connection to knowledge of grammatical morphology, word-finding abilities and verbal working memory. *International Journal of Language & Communication Disorders, 49*(4), 498–507. https://doi.org/10.1111/1460-6984.12091

Mather, G., & West, S. (1993). Recognition of animal locomotion from dynamic point-light displays. *Perception, 22*(7), 759–766.

McGugin, R. W., Gatenby, J. C., Gore, J. C., & Gauthier, I. (2012). High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance. *Proceedings of the National Academy of Sciences, 109*(42), 17063–17068. https://doi.org/10.1073/pnas.1116333109

Mitchell, P. (1997). *Introduction to theory of mind: Children, autism and apes*. London: Edward Arnold Publishers.

O'Connell, L. A., & Hofmann, H. A. (2011). The vertebrate mesolimbic reward system and social behavior network: A comparative synthesis. *The Journal of Comparative Neurology, 519*(18), 3599–3639. https://doi.org/10.1002/cne.22735

Ozonoff, S., Young, G. S., Landa, R. J., Brian, J., Bryson, S., Charman, T., … Iosif, A.-M. (2015). Diagnostic stability in young children at risk for autism spectrum disorder: A baby siblings research consortium study. *Journal of Child Psychology and Psychiatry, 56*(9), 988–998. https://doi.org/10.1111/jcpp.12421

Peterson, C., Slaughter, V., Moore, C., & Wellman, H. M. (2016). Peer social skills and theory of mind in children with autism, deafness, or typical development. *Developmental Psychology, 52*(1), 46–57. https://doi.org/10.1037/a0039833

Pierce, K., Marinero, S., Hazin, R., McKenna, B., Barnes, C. C., & Malige, A. (2016). Eye tracking reveals abnormal visual preference for geometric images as an early biomarker of an autism spectrum disorder subtype associated with increased symptom severity. *Biological Psychiatry, 79*(8), 657–666. https://doi.org/10.1016/j.biopsych.2015.03.032

Pierce, R. C., & Kumaresan, V. (2006). The mesolimbic dopamine system: The final common pathway for the reinforcing effect of drugs of abuse? *Neuroscience and Biobehavioral Reviews, 30*(2), 215–238. https://doi.org/10.1016/j.neubiorev.2005.04.016

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Reid, V. M., Dunn, K., Young, R. J., Amu, J., Donovan, T., & Reissland, N. (2017). The human fetus preferentially engages with face-like visual stimuli. *Current Biology, 27*(12), 1825. https://doi.org/10.1016/j.cub.2017.05.044

Sacco, R., Gabriele, S., & Persico, A. M. (2015). Head circumference and brain size in autism spectrum disorder: A systematic review and meta-analysis. *Psychiatry Research, 234*(2), 239–251. https://doi.org/10.1016/j.pscychresns.2015.08.016

Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology, 16*(2), 235–239. https://doi.org/10.1016/j.conb.2006.03.001

Schultz, R. T. (2005). Developmental deficits in social perception in autism: The role of the amygdala and fusiform face area. *International Journal of Developmental Neuroscience, 23*(2–3), 125–141.

Schultz, R. T., Grelotti, D. J., Klin, A., Kleinman, J., Van der Gaag, C., Marois, R., & Skudlarski, P. (2003). The role of the fusiform face area in social cognition: Implications for the pathobiology of autism. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 358*(1430), 415–427. https://doi.org/10.1098/rstb.2002.1208

Sedgewick, F., Hill, V., Yates, R., Pickering, L., & Pellicano, E. (2016). Gender differences in the social motivation and friendship experiences of autistic and non-autistic adolescents. *Journal of Autism and Developmental Disorders, 46*, 1297–1306. https://doi.org/10.1007/s10803-015-2669-1

Shah, A., & Frith, U. (1993). Why do autistic individuals show superior performance on the block design task? *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 34*(8), 1351–1364.

Shanok, N. A., Jones, N. A., & Lucas, N. N. (2019). The nature of facial emotion recognition impairments in children on the autism spectrum. *Child Psychiatry and Human Development, 50*, 661. https://doi.org/10.1007/s10578-019-00870-z

Shic, F., Macari, S., & Chawarska, K. (2014). Speech disturbs face scanning in 6-month-old infants who develop autism spectrum disorder. *Biological Psychiatry, 75*(3), 231–237. https://doi.org/10.1016/j.biopsych.2013.07.009

Supekar, K., Kochalka, J., Schaer, M., Wakeman, H., Qin, S., Padmanabhan, A., & Menon, V. (2018). Deficits in mesolimbic reward pathway underlie social interaction impairments in children with autism. *Brain, 141*, 2795. https://doi.org/10.1093/brain/awy191

Thompson, J. C., Clarke, M., Stewart, T., & Puce, A. (2005). Configural processing of biological motion in human superior temporal sulcus. *The Journal of Neuroscience, 25*(39), 9059–9066.

Thompson, J. C., Hardee, J. E., Panayiotou, A., Crewther, D., & Puce, A. (2007). Common and distinct brain activation to viewing dynamic sequences of face and hand movements. *NeuroImage, 37*(3), 966–973.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Wolf, J. M., Tanaka, J. W., Klaiman, C., Cockburn, J., Herlihy, L., Brown, C., … Schultz, R. T. (2008). Specific impairment of face-processing abilities in children with autism spectrum disorder using the let's face it! skills battery. *Autism Research, 1*(6), 329–340.

Zhang, F., Savadjiev, P., Cai, W., Song, Y., Rathi, Y., Tunç, B., … O'Donnell, L. J. (2018). Whole brain white matter connectivity analysis using machine learning: An application to autism. *NeuroImage, 172*, 826–837. https://doi.org/10.1016/j.neuroimage.2017.10.029

Zhu, X., Bhatt, R. S., & Joseph, J. E. (2016). Pruning or tuning? Maturational profiles of face specialization during typical development. *Brain and Behavior: A Cognitive Neuroscience Perspective, 6*(6), e00464. https://doi.org/10.1002/brb3.464

Self-Other Distinction

Tslil Simantov, Michael Lombardo, Simon Baron-Cohen, and Florina Uzefovsky

From sensation to emotion, we use our own experiences to better understand others. But, in order to truly understand and engage with others we must be able to comprehend that others, and their experiences, can be different to our own. We need to be able to draw the line between our own experiences and that of others to be able to comprehend the world around us more accurately. This chapter will review what is known regarding different aspects of self-other distinction, how it develops, and how this concept relates to autism spectrum conditions (henceforth, autism).

Self-Other Coupling

In biology, one of the defining principles of a living thing is that it is surrounded by a permeable membrane which separates it from its environment, but also allows for interaction with the environment (Ruiz-Mirazo, Peretó, & Moreno, 2004). That is, self-definition is crucial for the existence of the self. Yet, in order to understand others, much like in basic biology, the membrane must be permeable to allow for communication. Indeed, evolutionary and neuroscientific theories posit a mechanism of

T. Simantov · F. Uzefovsky (✉)
Psychology Department, Ben Gurion University in the Negev, Beer Sheva, Israel

Zlotowski Center for Neuroscience, Ben Gurion University in the Negev, Beer Sheva, Israel
e-mail: tslilsi@post.bgu.ac.il; florina@bgu.ac.il

M. Lombardo
Italian Institute of Technology, Trento, Italy

S. Baron-Cohen
Department of Psychiatry, Autism Research Centre, University of Cambridge, Cambridge, UK

CLASS Clinic, Cambridgeshire and Peterborough Mental Health NHS Foundation Trust, Cambridge, UK

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_5

self-other coupling that facilitates such communication. The perception-action evolutionary model (Preston & De Waal, 2002) suggests that understanding others is based primarily on a coupling between perception and action. In this case, between what we perceive the other doing, and a representation of doing the same thing, thus enabling an understanding of the meaning and goal of the other's actions. A myriad of neuroscience studies, from monkeys to humans, support the idea of a sensory-motor coupling in the brain (Gallese & Goldman, 1998; Iacoboni, 2009). Additional studies show similar activation patterns when observing another experience an emotion (e.g., pain) and feeling the same emotion (Shamay-Tsoory, 2011; Singer et al., 2004). Indeed, if all our experiences were identical we would have a perfect understating of each other's sensory, mental, and emotional states. But this is not the case. Our representations are shaped both by past experience and current perception. Thus, overreliance on our own representations can distort communication. Therefore, an additional mechanism is needed to assign representations to self versus other. This chapter focuses on this mechanism and the clinical implications of its breakdown.

Self-Identity

For one to be able to distinguish self from another, a rudimentary sense of self must exist. It is believed that a rudimentary, sensory sense of self develops through the experience of coupling between generating motor actions and the sensory experience these cause (Rochat & Striano, 2000). Self-identity is clearly present around the middle of the second year, once children can recognize themselves in the mirror (Bertenthal & Fischer, 1978), but rudimentary self-knowledge appears as early as at 2 months of age (Rochat & Striano, 2000), and there is some evidence for body self-knowledge (and self-other distinction) even in the womb (Castiello et al., 2010). According to Neisser's developmental theory (Neisser, 1991), self-knowledge develops through transactions with others (interpersonal self) and through transactions with the environment (ecological self), implying that both sensory and social input are important for the development of the self. Importantly, this also implies that the self develops through the distinction between self and other.

Self-Other Distinction

Self-other distinction can be defined as the ability to implicitly or explicitly differentiate between sensations, knowledge, and feelings of the self and the other. Thus, we can differentiate between sensory, cognitive, and emotional aspects of self-other distinction.

Sensory

Self-other distinction in the sensory domain is evident in the differential pattern of response to perceptions originating from the self and from another. For example, perceiving one's own face is processed faster than the perception of another familiar or unfamiliar face (Sui, Chechlacz, Rotshtein, & Humphreys, 2013; Tong & Nakayama, 1999). Similarly, newborns respond differently to the sound of their own crying as compared to the sound of another infant's cry (Dondi, Simion, & Caltran, 1999), reflecting a perceptual distinction between the two sounds. Another way of investigating self-other distinction in the sensory domain is through examining imitation. Previous research showed that copying the movement pattern of another is a dominant response (e.g., moving the same finger that another is moving; de Guzman, Bird, Banissy, & Catmur, 2016), which implies that the process of differentiation requires some effort. Indeed, studies in motor imitation suggest that the perception of the other strongly directs behavior (Brass & Heyes, 2005; Brass, Ruby, & Spengler, 2009; Santiesteban, White, et al., 2012; Wang & Hamilton, 2012). On the other hand, studies of visual perspective taking suggest that inhibiting one's own point of view in order to act based on that of the other's requires effort (Keysar, Lin, & Barr, 2003). This apparent contradiction may be explained by the methodology of these studies. Typically, in studies of imitation, participants are asked to look at the other, while in studies of visual perspective taking, the salient visual cue is that of the point of view of the self. Thus, it seems that the visual modality presents especially strong input regarding the relevance and importance of the stimuli. For both types of studies, self-other distinction requires some effort, suggesting that coupling is more dominant than differentiation. Moreover, even though our sense of bodily self is early developing and firmly established, it can also be manipulated.

Studies using body ownership illusions show that under certain conditions, the sense of self can be manipulated to include a fake or another's body part, and even whole body or face (Kilteni, Maselli, Kording, & Slater, 2015; Thirioux, Wehrmann, Langbour, Jaafari, & Berthoz, 2016). The most famous illusion is that of the rubber hand. In this case, a participant's hand is concealed from view, and she observes a rubber hand being touched, while simultaneously experiencing touch herself (on the concealed hand). This leads to the experience of the fake, rubber hand as her own. To conclude, the sensory experience of body ownership develops early and is the basis for any interaction with the surrounding physical and social environment. However, this process relies on constant sensory input and can therefore be manipulated, and requires effective control over the coupling between self and other.

Interestingly, this type of body ownership illusion has implications not only for the sensory, but also for the social domain. For example, a recent study found that inducing the illusionary perception of ownership of a hand was related to an increase in mu rhythm and beta rhythm desynchronization, which is thought to be an index of mirror system activity measured using electroencephalogram (EEG) (Riecansky, Lengersdorff, Pfabigan, & Lamm, 2020). Another study used a procedure in which the participant's cheek was brushed, while observing another's cheek being brushed synchronously. Not only did that induce an increase in the face ownership of the other, but also in the similarity between the perceived personality of the other and that of the participant's (Paladino, Mazzurega, Pavani, & Schubert, 2010). These studies show that the degree of sensory distinction between self and other is important not only for basic perception processes, but also has far-reaching implications for the social domain.

What is the mechanism by which sensory differentiation contributes to cognitive and emotional differentiation? One study provides some insight. de Guzman and colleagues trained participants either to enhance or reduce self-other differentiation by asking participants to either copy the finger movement completed by another or to execute an opposite finger movement (de Guzman et al., 2016). These two opposing training regimes had opposing effects both on imitation control (the ability to control a dominant response of motor imitation), and on corticospinal empathy for pain (motor evoked potentials in response to another's pain that are similar to those evoked when one experiences pain herself). In a second study, the training effects were also related to differences in self-reported overall empathy. Together with the previously reviewed studies, this study suggests that while the overlap in representation of self and other is specific per domain (e.g., visual, motor, pain), the mechanism by which differentiation occurs is non-domain specific.

Cognitive

Self-other distinction in the cognitive domain allows us to distinguish between our own knowledge and beliefs, and those of the other. This falls under the broader definition of Theory of Mind (ToM) (Baron-Cohen, 2000), a crucial ability for all aspects of social interaction, because it is central to the perception of another as an agent with distinct knowledge, beliefs, and desires. Neurobiological research from the last few decades points to two possible mechanisms or brain systems involved in understanding others. According to the simulation theory, we understand others by simulating their situation, and inferring what they feel or know based on the knowledge and feelings that arise (Gordon, 1992).

This theory has gained much support in the last decade due to the discovery of the mirror neurons in monkeys (motor neurons that are active both when a monkey preforms a certain action and when the monkey observes the same action), and is thought to underlie understanding of others' actions across domains—action goals, knowledge, desires, beliefs, and feelings (Gallese & Goldman, 1998). Indeed, based on findings regarding the mirror system in humans and animals, observing another's situation activates the same areas in the brain[1] as experiencing the same situation

[1] It is very difficult to conduct single cell recordings in humans and therefore it is unclear whether the activity is in fact in the same neurons or adjacent neurons within the same general brain area.

(Iacoboni, 2009; Keysers & Gazzola, 2018; Mukamel, Ekstrom, Kaplan, Iacoboni, & Fried, 2010; Santiesteban, White, et al., 2012). Clearly, this mechanism relies on a coupling in time and space (same or similar brain areas) between the experience of the other and one's own experience, thus requiring a mechanism for self-other distinction that will allow one to distinguish between a first-hand experience and simulation. The mirror neuron theory is however not without its critics (Call & Tomasello, 2008; Keysers & Gazzola, 2010). One key criticism is that non-human animals seem to have mirror neurons but fail theory of mind tasks. A second key criticism is that the regions containing mirror neurons in the brain seem to be expanding such that it may not be a very helpful concept.

The second view of the mechanism underlying ToM is rooted in the so-called theory theory. According to this view, rather than simulating the other's experience and using this to understand what they might think or feel, we understand others by perceiving their situation and *inferring* the other's knowledge, beliefs, and emotions (Wellman, 1990). Both theories suggest that our representations of self and other must be distinguished, but in the case of theory theory it is thought that we must inhibit our own knowledge, desires, and beliefs in order to allow for the representation of the other's knowledge, desires, and beliefs. Indeed, in this case as well, self-other distinction is necessary in order to represent the other's mind, and inhibit our own perspective. Developmentally, it is thought that ToM develops during early childhood until about 4 years of age, through a maturation of self-other distinction, even though recent research with infants suggests that ToM precursors appear already in the first 18 months of life (reviewed in Uzefovsky & Baron-Cohen, 2018), although it is unclear to what extent this reflects true self-other distinction. However, under certain conditions, even adults find it difficult to differentiate between their own knowledge and that of the other. This has been studied by Keyser and colleagues (Keysar et al., 2003), using the Director task. In this task, the participant is asked to move objects within a virtual cupboard, according to the instructions of a "director." The director is standing behind the cupboard, and his view of some of the shelves is obscured. The participant needs to take into account the difference in their visual perspective in order to accurately carry out the instructions of the director. Indeed, this is not an easy task, and adults often make mistakes in this task. Training in imitation inhibition increased the accuracy on this task (Santiesteban, White, et al., 2012), strengthening the interpretation that ToM relies on self-other differentiation, and that differentiation is not modality specific.

Affective

Insights into self-other distinction in the affective domain come from research on empathy. Empathy is defined as the ability to recognize and share in the emotions of others, while maintaining a self-other distinction (Uzefovsky & Knafo-Noam, 2017). Empathy is multifaceted, with both cognitive aspects (i.e., recognizing the other's mental states), and affective aspects (i.e., responding to the mental states of the other with an appropriate emotion). Both aspects require, paradoxically, that we use our own knowledge and emotions, as these are the bases for our understanding of the other, but also to maintain a clear understanding of which emotion is ours and which is of the other. At the most basic level, this effect is exemplified in the emotional egocentricity bias task (Silani, Lamm, Ruff, & Singer, 2013). During this task a participant receives either pleasant or unpleasant touch (e.g., being stroked by flower petals or holding some worms). This occurs while observing another's hand receiving either congruent or incongruent touch. The participant is asked to rate her own feelings and that of the other. In the case of incongruent trials, participants tend to rate the other's experience as more similar to their own (e.g., rating the pleasant touch received by the other as less pleasant when receiving unpleasant touch than when receiving pleasant touch). This bias suggests that we use our own experiences when judging that of the other.

In other cases, the other's emotions may impact our own. Neurobiological, evolutionary, and developmental theories of empathy suggest that self-other blurring, termed emotional contagion, occurs as the first step in empathy (Davidov, Zahn-Waxler, Roth-Hanania, & Knafo, 2013; Preston & De Waal, 2002). This means that perceiving another experiencing sadness creates a sensation of sadness within us as well. Thereafter, we can regulate our own emotions, in order to remain focused on the other, and use this initial empathic arousal to better understand the other and share in the other's emotions (Davidov et al., 2013). Emotion contagion is evident from birth, with neonates responding to another baby's cry by crying as well (Sagi & Hoffman, 1976). This response is attenuated with development, and with the development of emotion regulation, but ultimately can occur at any age, depending on the intensity of emotional arousal caused by the other's emotion (Davidov et al., 2013). Thus, emotion regulation, requiring some effort, is a critical part in changing the focus from self to other. If we are unable to regulate our own emotions, the distinction between self and other is blurred, and we experience the pain of the other as our own, an experience termed self-distress. Experiencing intense emotions and stress can inhibit our ability to distinguish between our own emotions and that of the other (Decety & Lamm, 2011). This effect can shed light on the centrality of self-other distinction, and the implications of failure to distinguish, on social behavior. Thus, when experiencing self-distress, focus is shifted from the other's emotional experience and needs, to our own. This means that our responses are geared towards relieving our own distress and not that of the other. Research by Batson's group and by developmental psychologists such as Zahn-Waxler and Eisenberg has shown that prosocial behavior is linked to feelings of empathy, while feelings of self-distress hinder such a response (Batson et al., 1988, 1989; Eisenberg et al., 1989; Eisenberg & Fabes, 1990; Young, Fox, & Zahn-Waxler, 1999).

Stress as a Modulator of Self-Other Distinction

As mentioned above, the experience of strong emotions can hinder self-other differentiation with consequences across domains. This is the mechanism hypothesized to underlie responses of personal distress to the distress of others (Davidov et al., 2013), shifting focus from other to self. A recent study directly examined the effects of stress on self-other differentiation (Tomova, von Dawans, Heinrichs, Silani, & Lamm, 2014). Stress was manipulated using the group version of the Trier Social Stress Test (TSST-G) (von Dawans, Kirschbaum, & Heinrichs, 2011), and salivary cortisol was used as a measure of experienced stress. Three measures of self-other differentiation were used; the first is a measure in the perception-motor domain, a measure of imitation inhibition (lifting the index or middle finger while observing another's hand preforming a congruent or incongruent movement). The second measure was the "Director task" described above (Keysar et al., 2003), and the third is a measure of emotional egocentricity bias (Silani et al., 2013), in which the participant receives either pleasant or unpleasant touch, while observing another's hand receiving congruent or incongruent touch. In the incongruent conditions, participants rate the other's experience as more similar to their own experience, thus manifesting bias in their judgments regarding the other's experiences. Across all measures of self-other distinction, stress induced lower self-other differentiation for men, and higher self-other differentiation for women (Tomova et al., 2014). This intriguing finding suggests that sex (gender) is related differentially to self-other distinction. This finding is in line with the tend-and-befriend theory of stress response (Taylor, 2006), according to which women respond to stress with a drive towards affiliation while men tend to respond with fight or flight, i.e., disengagement.

The Neurobiology of Self-Other Distinction

The salience of coupling versus distinction is also evident in the brain. Convincing evidence from research on the animal and human mirror system, using brain imaging methods such as functional Magnetic Resonance Imaging (fMRI) and single cells recordings, show that similar areas are active when one experiences a certain sensation or emotion as well as one perceives another experiencing the same (Gallese, 2005; Gallese, Fadiga, Fogassi, & Rizzolatti, 1996). Single cell recordings in mice and monkeys show the existence of sensory-motor neurons, and recent work by Keysers and Gazzola's group shows similar evidence for emotional mirror neurons in the rat's anterior cingulate cortex (ACC), responding specifically to pain in self and other (Carrillo et al., 2019). This study suggests that the coupling between self and other occurs at the level of the single neuron.

Which areas in the brain are responsible for distinguishing self and other? Two main brain areas, with possibly different roles, are thought to be involved. The first is the medial prefrontal cortex (MPFC). One of the most consistent findings in neuroimaging is that the MPFC, and in particular the ventromedial PFC (vMPFC) is activated during processing of information regarding the self (Amodio & Frith, 2006). This is true for many self-referential processes, including autobiographical memories, free-form thoughts about the self, introspection, judging one's own emotional response and other tasks (reviewed in Heatherton, 2011; Kelley et al., 2002; Ochsner et al., 2004). Thus, the vMPFC responds preferentially to any information

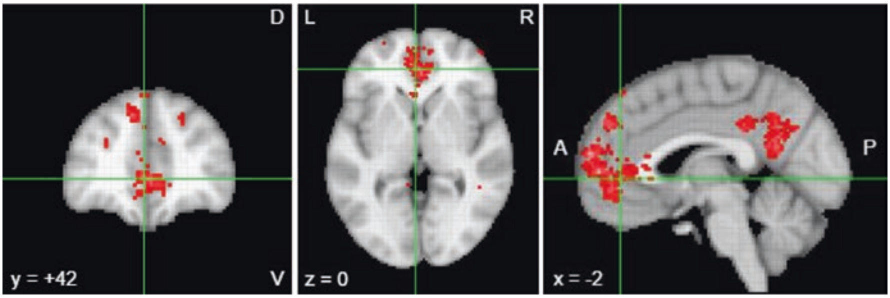


Fig. 1 Results of an automated meta-analysis conducted using NeuroSynth with respect to the term "self-referential"

or processing concerning the self, and as such, this signal can differentiate between self and other. Indeed, several meta-analyses of studies that compared self to other processing in typical populations, found effects within the vMPFC (Denny, Kober, Wager, & Ochsner, 2012; Lombardo et al., 2010). More specifically, findings suggest that several areas within the MPFC are more active for self-judgments. For the purpose of this chapter, we conducted an analysis using NeuroSynth (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011) with regard to the term "self-referential." Based on 166 studies mentioning this term, the majority of brain activity was localized to the vMPFC (See Fig. 1. Original analysis conducted on 16 July 2019. http://neurosynth.org/analyses/terms/self%20referential/).

The second major brain area implicated in self-other distinction is the right parietal cortex (Decety & Sommerville, 2003), and specifically, the right temporoparietal junction (rTPJ) and the right supramarginal gyrus (rSMG). Recent studies suggest that the rTPJ is involved in distinction in the cognitive domain and the rSMG is involved in distinction in the affective domain (Steinbeis, 2016). Several studies support this distinctive role for the two parietal areas. For example, a transcranial stimulation study showed that excitatory stimulation of the rTPJ increased the ability for self-other differentiation in visual perspective taking (Director's task), but had no effect on social judgments, albeit these did not require on-line self-other distinction (Santiesteban, Banissy, Catmur, & Bird, 2012). Indeed, a meta-analysis of studies comparing self and other judgments showed that the TPJ was more active for "other" judgments (Denny et al., 2012). While no distinction was made in the meta-analysis between the cognitive and affective domains, the majority of the studies included in the meta-analysis investigated the cognitive domain. On the other hand, the rSMG was found to be specifically involved in the emotional egocentricity bias task described above (Silani et al., 2013). In this study, disruption in the activity if the rSMG achieved using repeated transcranial magnetic stimulation, resulted in increased egocentricity bias. The distinct role of two brain areas for different domains suggests that the two domains may be independent of each other.

However, findings of one study by de Guzman and colleagues (2016) challenge this. In this study, participants were trained to either increase or decrease self-other differentiation in the motor domain, and then tested on an implicit empathy for pain measure (reduction in motor evoked potential, MEP, when observing another receiving a painful stimulation); Experiment (1) and an explicit empathy measure—the Questionnaire of Cognitive and Affective Empathy (QCAE) (Reniers, Corcoran, Drake, Shryane, & Völlm, 2011); Experiment (2). Those participants who received a one-session training to increase self-other motor distinction showed lower MEP when observing painful stimulation of another's hand as compared to touch. This reflects an increased empathy to the other's pain. Indeed, performance on the self-other motor distinction task was correlated with corticospinal empathy. Experiment 2 showed a similar change (although very small) on the questionnaire measure of empathy (de Guzman et al., 2016). These findings suggest that although there are two brain areas responsible for different domains of self-other distinction, the rTPJ and the rSMG, they are interconnected, and modifying the activity of one may influence the activity of the other.

The Role of Oxytocin

Oxytocin is an evolutionarily conserved neuromodulator of social behavior (Carter, 2014). Many studies suggest a central role for oxytocin in modulating social behavior, but the exact mechanism of action has been debated. The most empirically sound theory suggests that oxytocin acts to make social stimuli more salient (Shamay-Tsoory & Abu-Akel, 2016). Additional studies suggest that this may also involve effects on self-other distinction, but the findings are mixed. Some studies find a decrease in self bias or an increase in self-other blurring (Ruissen & de Bruijn, 2015; Zhao et al., 2016), while other studies show an increase in the focus on the other and a better ability to distinguish self from other (Abu-Akel, Palgi, Klein, Decety, & Shamay-Tsoory, 2015; Colonnello, Chen, Panksepp, & Heinrichs, 2013). Another recent study by Pfundmair and collegues (2018) investigated this discrepancy, hypothesizing that OT diminishes self-other distinction only at the implicit, but not the explicit level. In their studies, they used eye tracking and measures of mimicry to test for implicit self-other distinction and found that when using implicit measures intranasal administration typically was associated with diminished self-other distinction, but not so for explicit measures of self-other distinction (Pfundmair, Rimpel, Duffy, & Zwarg, 2018). Taken together, the findings suggest that further investigation into the role of oxytocin in self-other distinction is needed.

The above reviewed findings point to the crucial role of self-other distinction for social functioning. Therefore, it is not surprising that autism has been connected with difficulties in self-other distinction. In the next part, we will review findings relating to self-other distinction in autism.

Manifestations of Impaired Self-Other Distinction in Autism

Autism is a prevalent (Baio et al., 2018; Lai, Lombardo, & Baron-Cohen, 2014) and pervasive developmental condition characterized by difficulties in the social domain and in restrictive and repetitive interests and behaviors (American Psychiatric Association, 2013). The term "autism" emphasizes the difficulties in the self-to-other relationship. The term is derived from the Greek word "autos", which means "self." Psychiatrist Leo Kanner was one of the first to describe the condition and the tendency of children diagnosed as such to remain solitary (Kanner, 1943). Kanner described the children as happiest when left alone, indifferent to the presence of relatives and behaving as if people did not matter or exist. It seemed like the children Kanner described lived within themselves.

In the best case scenario, autism is diagnosed at an early age, around toddlerhood, primarily based on social difficulties. Early indicators include deficits in the emergence of joint attention and pretend play, deficits in reciprocal affective behavior, reduced response to own name, reduced imitation, delayed communication (both verbal and nonverbal), repetitive behaviors, sensory hypersensitivity, and motor delay (Lai et al., 2014). Furthermore, autistic individuals also show impaired reciprocity skills, regardless of cognitive or language ability (Carter, Davis, Klin, & Volkmar, 2005).

Self-Identity in Autism

The sense of self seems to be lacking in children diagnosed with autism from an early stage. Some studies show that autistic 1-year-old infants demonstrate less orienting to own name compared to typically developing infants (Osterling, Dawson, & Munson, 2002; Zwaigenbaum et al., 2005). Other studies showed that most autistic children at the ages of 3.5–12.7 years old succeed at the test of mirror self-recognition (Dawson & McKissick, 1984; Ferrari & Matthews, 1983; Neuman & Hill, 1978; Spiker & Ricks, 1984). In addition, studies show that autistic individuals are able to become perceptually aware of physical aspects of themselves (David et al., 2008; Williams & Happé, 2009a), but have diminished primary awareness of psychological aspects of self (Hobson, 1990). This suggests a difficulty in the social but not physical or perceptual aspects of self-identity.

The Autistic Self

Prior to understanding self-other distinction in autism, we first need to address questions regarding the structure of the autistic self and self-referential cognition, i.e., the ability and tendency to think of oneself as an agent. The self-reference effect is a tendency of people to encode information differently depending on how much they are implicated in the information. This effect results in that information regarding the self is better-processed (Symons & Johnson, 1997). Since the initial description of autism, this condition has been characterized by extreme egocentrism, with children described as being locked in "a world of their own" or surrounded by "a glass bubble" (Lombardo & Baron-Cohen, 2010). Therefore, a paradox arises—how can autism be characterized both by impaired in self-referential cognition, and extreme egocentrism? Lombardo and Baron-Cohen (2010) settle the paradox by dismissing the assumption that egocentrism and self-referential cognition are two independent phenomena. Both phenomena revolve around a common mechanism, a deficit within the neural circuitry coding for self-representations.

Autism and Self-Other Distinction

Difficulties in self-other distinction manifest in autism across the sensory, cognitive, and affective domains, suggesting that this is a core deficit in autism. We will review evidence for this below.

Sensory

Atypical sensory processing, across multiple modalities, is extremely prevalent in autism (Marco, Hinkley, Hill, & Nagarajan, 2011). There is vast literature indicating that individuals with autism do not readily imitate the actions of others (Rogers, 1999; Rogers & Pennington, 1991; Smith & Bryson, 1994). What is the basis for this observed deficit? One line of inquiry suggests that it is due to a specific deficit in motor imitation. Another line of research suggests that the deficits are in the basic ability to map the actions of others in order to imitate them correctly, especially when such actions are complex (Perner, 1996; Rogers, Bennetto, McEvoy, & Pennington, 1996; Smith & Bryson, 1994; Whiten & Brown, 1998). Conversely, a more recent study using fMRI found that autistic participants showed a significant hyperimitation of actions, including echolalia (involuntary repetition of another person's vocalizations) and echopraxia (involuntary repetition of another person's actions), and showed problems controlling the mirror system functions. As a result, autistic individuals exhibited a higher interference effect, meaning experienced more deficits in inhibiting automatic imitation compared with a matched control (Spengler, Bird, & Brass, 2010). How can these seemingly contradictory effects co-exist? It has been suggested that these phenomena arises due to difficulty in switching between the self and other (Bird & Viding, 2014) and we will expand on this below.

A recent study showed that when an observed touch is incongruent with a felt touch, high functioning autistic adults show deficits in signaling and EEG responses

(Deschrijver, Wiersema, & Brass, 2016). The task reflected self-other distinction processes, and therefore it was suggested that autism is associated with difficulties in distinguishing self from other based on touch. Other studies demonstrated how effects of the rubber hand illusion vary along the non-clinical to clinical autism spectrum (Cascio, Foss-Feig, Burnette, Heacock, & Cosby, 2012; Palmer, Paton, Hohwy, & Enticott, 2013; Paton, Hohwy, & Enticott, 2012). Altogether, it seems that autistic individuals experience difficulties in distinguishing the self from others on a sensorimotor level.

Another relevant study aimed at understanding the imitation of "style," meaning the qualities of a person's actions that were incidental to the accomplishment of a goal (Hobson & Lee, 1999). Findings showed that fewer participants with autism imitated the style of the demonstrator's actions, as compared to typically developing individuals. The authors inferred that there may be specific aspects of imitation that are abnormal in autism. Imitating style often reflects the intention of the person being imitated so this difference may reflect a difficulty with "theory of mind." Throughout typical development, these aspects of imitation may exert a significant contribution to establishing intersubjective contact, as compared to imitation of goal-directed actions.

Cognitive

Autistic individuals often find it hard to pass theory of mind (ToM) tasks (Baron-Cohen, Leslie, & Frith, 1985; Happé, 1995; Yirmiya, Erel, Shaked, & Solomonica-Levi, 1998). This is especially salient in the false-belief task. The false-belief task aims to test one's ability to differentiate between the situation in the real world, and another's (false) perception of the same situation. Baron-Cohen et al. (1985) showed that children with autism often fail or are extremely developmentally delayed in passing this task. Many following studies have reached the same conclusion, suggesting that autism is characterized by difficulties in conceptualizing mental states, and thus failing to attribute (false) beliefs to others (Baron-Cohen, Tager-Flusberg, & Cohen, 1994). Many studies thereafter showed that ToM impairments in individuals with autism are not due to a general deficit in meta-representation (i.e., the ability to represent an agent's many possible mental attitudes or interpretations of a statement or proposition), or in the broadly defined social domain (Williams & Happé, 2009b), but rather a specific deficit in the self-referential domain (Charman & Baron-Cohen, 1992). Frith and Happé (1999) argued that observed deficits in the social domain, and in meta-representation stem from deficits in the self-referential domain. They argue that if the mechanism that underlies the computation of mental states is dysfunctional, then self-knowledge is likely to be impaired in addition to the knowledge of other's minds. According to this theory, individuals with autism lack the cognitive machinery to represent their thoughts and feelings as thoughts and feelings and may know as little about their own minds as about the minds of other people.

Findings regarding ToM abilities in adulthood are mixed. One study showed that even high functioning adults with autism had impaired performance on theory of mind tasks compared to age-matched controls (Baron-Cohen et al., 2015; Baron-Cohen, Jolliffe, Mortimore, & Robertson, 1997). On the other hand, a different study showed that adults with high functioning autism and Asperger syndrome (who are now considered to be part of the autism spectrum) tend to develop better social understanding as they grow older, and show clear evidence of passing simple theory of mind tests (Bowler, 1992).

Nevertheless, the majority of findings point to a pervasive difficulty in the representations of mental states. This led to the formulation of "mindblindness" theory of autism, which provides a cognitive explanation for the social-communicative difficulties in autism (Baron-Cohen, 1997; Frith, 2001). According to this theory, individuals with autism find it difficult to attribute mental states to both self and other, as opposed to the general population, who possesses an intact mechanism for representing or attributing mental states to both self and other. Understanding others relies on perceiving the similarities and differences between self and other, an ability that is associated with self-other distinction. Thus, it is possible that some of the complexities in ascribing mental states to others build upon difficulties in self-other distinction.

Affective

As mentioned above, empathy consists of cognitive and emotional components, and both rely on the ability to distinguish between one's own and other's emotions and perceptions (Gonzalez-Liencres, Shamay-Tsoory, & Brüne, 2013). Autistic individuals are typically considered to be lacking in empathy (Baron-Cohen & Wheelwright, 2004), and more specifically cognitive empathy, while emotional empathy is considered intact (Baron-Cohen, 1997; Smith, 2006) or too intense, creating an experience of personal distress (Attwood, 2003; Rogers, Dziobek, Hassenstab, Wolf, & Convit, 2007). Inadequate integration/differentiation between self-other in autism may also explain the empathic imbalance observed in autism (Richer, 2001). According to this theory, without the cognitive ability to understand two different perspectives simultaneously, the emotions of people with autism may become entangled with the emotions of the other. One explanation for this deficit is an inadequate control over the self to other "switch" (Bird & Viding, 2014). The self/other switch has two functions—adjusting the information processing system so that one's own affective state is appropriate to the other person's state, and then to tag that the empathizer's current state is appropriate for the other. According to the theory, the default state of the self/other switch is "self," and that switching to "other" is an active process. Therefore, one of the functions of the self-other switch is to direct attention towards others, and individuals with autism exhibit less social attention (Dawson, Webb, & McPartland, 2005). According to Bird and Viding (2014), autistic individuals would be less likely to draw a sharp distinction between the self and other and therefore may be more affected by another's state, compared to typically developing individuals. Supporting this view are several studies that show that individuals with autism often report greater personal distress as compared to neurotypical controls, that is a feelings of distress caused by the distress of another being experienced as one's own (Attwood, 2003; Rogers et al., 2007).

A recent study aimed to understand whether empathic responding in autistic adults is modulated by control over self-other distinction. This was done through comparing participants' reactions to observing another's pain after being imitated, as compared to no imitation. Researchers hypothesized that if empathic responding in autistic individuals is generally decreased or increased, and not susceptible to self-other manipulations, the imitation manipulation will not influence the empathy response for pain. Findings suggest that high functioning autistic adults and typically developed adults show equal, increased affective responses when observing another in pain, irrespective of being imitated or not. The findings indicate that there is no general empathy deficit in autism (De Coster, Wiersema, Deschrijver, & Brass, 2018). Compatible with this research, it is important to note that Bird and Viding (2014), and more recently de Guzman et al. (2016), argue that adequate empathic responding is specifically dependent upon *control* over self-other representations, rather than self-other representations themselves.

Neurobiology of Self-Other Distinction in Autism

Many of the difficulties and differences mentioned above are associated with underlying neural mechanisms. Mirror neuron system dysfunction is assumed to be one of the main deficits in autism, and to contribute to a self-other matching deficit (Williams, Whiten, Suddendorf, & Perrett, 2001). Self-other matching ability involves forming and coordinating mental representations of the self and others. Studies with humans and monkeys showed that mirror neurons can selectively respond to specific intentions, indicating that they are involved in the internal representations of another (Fogassi et al., 2005; Iacoboni et al., 2005). As such, mirror neuron function has been linked to imitation (Iacoboni et al., 2001; Nishitani & Hari, 2000), and theoretically linked to theory of mind (Gallese & Goldman, 1998). As described, both of these processes are related to self-other distinction and are impaired among autistic individuals (Baron-Cohen et al., 1985; Happé, 1995; Rogers, 1999; Rogers & Pennington, 1991; Smith & Bryson, 1994; Yirmiya et al., 1998). Note that this "broken mirror" theory of autism has been criticized on the grounds that autistic people can imitate while they struggle with ToM/cognitive empathy (Hamilton, Brindley, & Frith, 2007).

One interesting study investigated the functioning of neural systems involved in shared representations for self and others in autistic children. Using a self-face recognition paradigm, the researchers found that children diagnosed with autism exhibit decreased neural responses to viewing faces of others compared to viewing faces of themselves, as compared to controls (Uddin et al., 2008). In fact, the study showed that children with autism do not activate shared regions for self- and other-face processing, which points to functional dissociation between the representation of self versus others, to neural characteristics of self-focus and decreased social understanding.

Another study among autistic and typically developed adults observed specific disruptions in the neural systems involved in preferentially coding for self-information. The study showed that, in the autism group, the middle cingulate cortex responded more to other-mentalizing, rather than preferentially responding to self-mentalizing (Lombardo et al., 2010). Among others, the cingulate cortex is responsible for selective attention (Gabriel, Burhans, Talk, & Scalf, 2002). The reversed responding of the middle cingulate cortex might imply that an altered pattern of selective attention to social cues is connected to difficulties in self-other distinction. Furthermore, the study also showed a complete lack of preferential responsiveness to self-information in the vMPFC of individuals with autism, meaning their vMPFC responded to self and other judgments equivalently (Lombardo et al., 2010). In addition, in autistic individuals there was a reduced functional connectivity between vMPFC and other areas associated with lower level embodied representations, such as ventral premotor and somatosensory cortex. Interestingly, individuals whose vMPFC showed the largest distinction between mentalizing about self and other were least socially impaired in early childhood. Interestingly, a recent study showed that this effect was observed primarily in males, and not females, and for those females who showed more female-neurotypical patterns of activation in the vMPFC during self-representation were also higher on camouflaging (acting as behaviorally neurotypical), suggesting that self-other distinction is important for the compensatory techniques of camouflaging (Lai et al., 2019). These neurobiological findings lend further support to the notion of difficulties in self-other distinction in autism.

Oxytocin

One recent study investigated the role of oxytocin receptor (*OXTR*) single nucleotide polymorphisms (SNPs) in emotion recognition in adolescents with autism. The findings showed that genotype had a direct association, as well as interacted with diagnosis to predict activity within the right supramarginal gyrus (rSMG). Taken together with the findings connecting rSMG activity with self-other distinction within the affective domain, the findings have been interpreted as supporting the role of rSMG and the oxytocin system in modulating self-other distinction across the spectrum from neurotypical to autistic (Uzefovsky et al., 2019).

Summary

Taken together, the findings reviewed above suggest that self-other differentiation occurs across several processing domains—sensory, cognitive, and affective. This very basic process emphasizes that our brains are wired for connecting with others, but also for maintaining a stable sense of self. Only when this process of sharing and disengagement happens accurately and smoothly are we able to genuinely and accurately connect with others. When the switch fails as in the case of increased stress or autism, it becomes very difficult to flexibly interact with others' mental states.

References

Abu-Akel, A., Palgi, S., Klein, E., Decety, J., & Shamay-Tsoory, S. (2015). Oxytocin increases empathy to pain when adopting the other-but not the self-perspective. *Social Neuroscience, 10*(1), 7–15.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. Washington, DC: American Psychiatric Association Publishing.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*(4), 268.

Attwood, T. (2003). *Why does Chris do that?: Some suggestions regarding the cause and management of the unusual behaviour of children and adults with autism and Asperger syndrome*. Shawnee, KS: AAPC Publishing.

Baio, J., Wiggins, L., Christensen, D. L., Maenner, M. J., Daniels, J., Warren, Z., … White, T. (2018). Prevalence of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveillance Summaries, 67*(6), 1.

Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT press.

Baron-Cohen, S. (2000). Theory of mind and autism: A review. In *International review of research in mental retardation* (Vol. 23, pp. 169–184). Amsterdam: Elsevier.

Baron-Cohen, S., Bowen, D. C., Holt, R. J., Allison, C., Auyeung, B., Lombardo, M. V., … Lai, M.-C. (2015). The "reading the mind in the eyes" test: Complete absence of typical sex difference in ~400 men and women with autism. *PLoS One, 10*(8), e0136521.

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child Psychology and Psychiatry, 38*(7), 813–822.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*(1), 37–46.

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders, 34*(2), 163–175.

Baron-Cohen, S. E., Tager-Flusberg, H. E., & Cohen, D. J. (1994). *Understanding other minds: Perspectives from autism*. Paper presented at the Most of the chapters in this book were presented in draft form at a workshop in Seattle, 1991.

Batson, C. D., Batson, J. G., Griffitt, C. A., Barrientos, S., Brandt, J. R., Sprengelmeyer, P., & Bayly, M. J. (1989). Negative-state relief and the empathy—Altruism hypothesis. *Journal of Personality and Social Psychology, 56*(6), 922.

Batson, C. D., Dyck, J. L., Brandt, J. R., Batson, J. G., Powell, A. L., McMaster, M. R., & Griffitt, C. (1988). Five studies testing two new egoistic alternatives to the empathy-altruism hypothesis. *Journal of Personality and Social Psychology, 55*(1), 52.

Bertenthal, B. I., & Fischer, K. W. (1978). Development of self-recognition in the infant. *Developmental Psychology, 14*(1), 44.

Bird, G., & Viding, E. (2014). The self to other model of empathy: Providing a new framework for understanding empathy impairments in psychopathy, autism, and alexithymia. *Neuroscience & Biobehavioral Reviews, 47*, 520–532.

Bowler, D. M. (1992). "Theory of Mind" in Asperger's syndrome Dermot M. Bowler. *Journal of Child Psychology and Psychiatry, 33*(5), 877–893.

Brass, M., & Heyes, C. (2005). Imitation: Is cognitive neuroscience solving the correspondence problem? *Trends in Cognitive Sciences, 9*(10), 489–495.

Brass, M., Ruby, P., & Spengler, S. (2009). Inhibition of imitative behaviour and social cognition. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 364*(1528), 2359–2367.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences, 12*(5), 187–192.

Carrillo, M., Han, Y., Migliorati, F., Liu, M., Gazzola, V., & Keysers, C. (2019). Emotional mirror neurons in the rat's anterior cingulate cortex. *Current Biology, 29*, 1301.

Carter, A. S., Davis, N. O., Klin, A., & Volkmar, F. R. (2005). Social development in autism. *Handbook of Autism and Pervasive Developmental Disorders, 1*, 312–334.

Carter, C. S. (2014). Oxytocin pathways and the evolution of human behavior. *Annual Review of Psychology, 65*, 17–39.

Cascio, C. J., Foss-Feig, J. H., Burnette, C. P., Heacock, J. L., & Cosby, A. A. (2012). The rubber hand illusion in children with autism spectrum disorders: Delayed influence of combined tactile and visual input on proprioception. *Autism, 16*(4), 406–419.

Castiello, U., Becchio, C., Zoia, S., Nelini, C., Sartori, L., Blason, L., … Gallese, V. (2010). Wired to be social: The ontogeny of human interaction. *PLoS One, 5*(10), e13199. https://doi.org/10.1371/journal.pone.0013199

Charman, T., & Baron-Cohen, S. (1992). Understanding drawings and beliefs: A further test of the meta representation theory of autism: A research note. *Journal of Child Psychology and Psychiatry, 33*(6), 1105–1112.

Colonnello, V., Chen, F. S., Panksepp, J., & Heinrichs, M. (2013). Oxytocin sharpens self-other perceptual boundary. *Psychoneuroendocrinology, 38*(12), 2996–3002.

David, N., Gawronski, A., Santos, N. S., Huff, W., Lehnhardt, F.-G., Newen, A., & Vogeley, K. (2008). Dissociation between key processes of social cognition in autism: Impaired mentalizing but intact sense of agency. *Journal of Autism and Developmental Disorders, 38*(4), 593–605.

Davidov, M., Zahn-Waxler, C., Roth-Hanania, R., & Knafo, A. (2013). Concern for others in the first year of life: Theory, evidence, and avenues for research. *Child Development Perspectives, 7*, 126–131. https://doi.org/10.1111/cdep.12028

Dawson, G., & McKissick, F. C. (1984). Self-recognition in autistic children. *Journal of Autism and Developmental Disorders, 14*(4), 383–394.

Dawson, G., Webb, S. J., & McPartland, J. (2005). Understanding the nature of face processing impairment in autism: Insights from behavioral and electrophysiological studies. *Developmental Neuropsychology, 27*(3), 403–424.

De Coster, L., Wiersema, J. R., Deschrijver, E., & Brass, M. (2018). The effect of being imitated on empathy for pain in adults with high-functioning autism: Disturbed self–other distinction leads to altered empathic responding. *Autism, 22*(6), 712–727.

de Guzman, M., Bird, G., Banissy, M. J., & Catmur, C. (2016). Self-other control processes in social cognition: From imitation to empathy. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 371*(1686), 20150079. https://doi.org/10.1098/rstb.2015.0079

Decety, J., & Lamm, C. (2011). 15 Empathy versus personal distress: Recent evidence from social neuroscience. In *The social neuroscience of empathy* (pp. 199–213). Cambridge, MA: MIT Press.

Decety, J., & Sommerville, J. A. (2003). Shared representations between self and other: A social cognitive neuroscience view. *Trends in Cognitive Sciences, 7*(12), 527–533. https://doi.org/10.1016/j.tics.2003.10.004

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self-and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience, 24*(8), 1742–1752.

Deschrijver, E., Wiersema, J. R., & Brass, M. (2016). Action-based touch observation in adults with high functioning autism: Can compromised self-other distinction abilities link social and sensory everyday problems? *Social Cognitive and Affective Neuroscience, 12*(2), 273–282.

Dondi, M., Simion, F., & Caltran, G. (1999). Can newborns discriminate between their own cry and the cry of another newborn infant? *Developmental Psychology, 35*(2), 418.

Eisenberg, N., & Fabes, R. A. (1990). Empathy: Conceptualization, measurement, and relation to prosocial behavior. *Motivation and Emotion, 14*(2), 131–149.

Eisenberg, N., Fabes, R. A., Miller, P. A., Fultz, J., Shell, R., Mathy, R. M., & Reno, R. R. (1989). Relation of sympathy and personal distress to prosocial behavior: A multimethod study. *Journal of Personality and Social Psychology, 57*(1), 55.

Ferrari, M., & Matthews, W. S. (1983). Self-recognition deficits in autism: Syndrome-specific or general developmental delay? *Journal of Autism and Developmental Disorders, 13*(3), 317–324.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: From action organization to intention understanding. *Science, 308*(5722), 662–667.

Frith, U. (2001). Mind blindness and the brain in autism. *Neuron, 32*(6), 969–979.

Frith, U., & Happé, F. (1999). Theory of mind and self-consciousness: What is it like to be autistic? *Mind & Language, 14*(1), 82–89.

Gabriel, M., Burhans, L., Talk, A., & Scalf, P. (2002). Cingulate cortex. *Encyclopedia of the Human Brain, 1*, 775–791.

Gallese, V. (2005). Embodied simulation: From neurons to phenomenal experience. *Phenomenology and the Cognitive Sciences, 4*(1), 23–48.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119*(2), 593–609.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2*(12), 493–501.

Gonzalez-Liencres, C., Shamay-Tsoory, S. G., & Brüne, M. (2013). Towards a neuroscience of empathy: Ontogeny, phylogeny, brain mechanisms, context and psychopathology. *Neuroscience & Biobehavioral Reviews, 37*(8), 1537–1548.

Gordon, R. M. (1992). The simulation theory: Objections and misconceptions. *Mind & Language, 7*(1-2), 11–34.

Hamilton, A. F. C., Brindley, R. M., & Frith, U. (2007). Imitation and action understanding in autistic spectrum disorders: How valid is the hypothesis of a deficit in the mirror neuron system? *Neuropsychologia, 45*(8), 1859–1868.

Happé, F. G. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development, 66*(3), 843–855.

Heatherton, T. F. (2011). Neuroscience of self and self-regulation. *Annual Review of Psychology, 62*, 363–390.

Hobson, R. P. (1990). On the origins of self and the case of autism. *Development and Psychopathology, 2*(2), 163–181.

Hobson, R. P., & Lee, A. (1999). Imitation and identification in autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines, 40*(4), 649–659.

Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annual Review of Psychology, 60*(1), 653–670. https://doi.org/10.1146/annurev.psych.60.110707.163604

Iacoboni, M., Koski, L. M., Brass, M., Bekkering, H., Woods, R. P., Dubeau, M.-C., … Rizzolatti, G. (2001). Reafferent copies of imitated actions in the right superior temporal cortex. *Proceedings of the National Academy of Sciences, 98*(24), 13995–13999.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology, 3*(3), e79.

Kanner, L. (1943). Autistic disturbances of affective contact. *The Nervous Child, 2*(3), 217–250.

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience, 14*(5), 785–794.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*(1), 25–41. https://doi.org/10.1016/S0010-0277(03)00064-7

Keysers, C., & Gazzola, V. (2010). Social neuroscience: Mirror neurons recorded in humans. *Current Biology, 20*(8), R353–R354.

Keysers, C., & Gazzola, V. (2018). Chapter 4 - Neural correlates of empathy in humans, and the need for animal models. In K. Z. Meyza & E. Knapska (Eds.), *Neuronal correlates of empathy* (pp. 37–52). New York, NY: Academic Press.

Kilteni, K., Maselli, A., Kording, K. P., & Slater, M. (2015). Over my fake body: Body ownership illusions for studying the multisensory basis of own-body perception. *Frontiers in Human Neuroscience, 9*, 141.

Lai, M., Lombardo, M., & Baron-Cohen, S. (2014). Autism. *Lancet, 383*(9920), 896–910.

Lai, M.-C., Lombardo, M. V., Chakrabarti, B., Ruigrok, A. N., Bullmore, E. T., Suckling, J., … Baron-Cohen, S. (2019). Neural self-representation in autistic women and association with 'compensatory camouflaging'. *Autism, 23*(5), 1210–1223.

Lombardo, M. V., & Baron-Cohen, S. (2010). Unraveling the paradox of the autistic self. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(3), 393–403.

Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Sadek, S. A., Pasco, G., Wheelwright, S. J., … Baron-Cohen, S. (2010). Atypical neural self-representation in autism. *Brain, 133*(2), 611–624. https://doi.org/10.1093/brain/awp306

Marco, E. J., Hinkley, L. B., Hill, S. S., & Nagarajan, S. S. (2011). Sensory processing in autism: A review of neurophysiologic findings. *Pediatric Research, 69*(5 Pt 2), 48R–54R.

Mukamel, R., Ekstrom, A. D., Kaplan, J., Iacoboni, M., & Fried, I. (2010). Single-neuron responses in humans during execution and observation of actions. *Current Biology, 20*(8), 750–756. https://doi.org/10.1016/j.cub.2010.02.045

Neisser, U. (1991). Two perceptually given aspects of the self and their development. *Developmental Review, 11*(3), 197–209.

Neuman, C. J., & Hill, S. D. (1978). Self-recognition and stimulus preference in autistic children. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology, 11*(6), 571–578.

Nishitani, N., & Hari, R. (2000). Temporal dynamics of cortical representation for action. *Proceedings of the National Academy of Sciences, 97*(2), 913–918.

Ochsner, K. N., Knierim, K., Ludlow, D. H., Hanelin, J., Ramachandran, T., Glover, G., & Mackey, S. C. (2004). Reflecting upon feelings: An fMRI study of neural systems supporting the attribution of emotion to self and other. *Journal of Cognitive Neuroscience, 16*(10), 1746–1772.

Osterling, J. A., Dawson, G., & Munson, J. A. (2002). Early recognition of 1-year-old infants with autism spectrum disorder versus mental retardation. *Development and Psychopathology, 14*(2), 239–251.

Paladino, M.-P., Mazzurega, M., Pavani, F., & Schubert, T. W. (2010). Synchronous multisensory stimulation blurs self-other boundaries. *Psychological Science, 21*(9), 1202–1207. https://doi.org/10.1177/0956797610379234

Palmer, C. J., Paton, B., Hohwy, J., & Enticott, P. G. (2013). Movement under uncertainty: The effects of the rubber-hand illusion vary along the nonclinical autism spectrum. *Neuropsychologia, 51*(10), 1942–1951.

Paton, B., Hohwy, J., & Enticott, P. G. (2012). The rubber hand illusion reveals proprioceptive and sensorimotor differences in autism spectrum disorders. *Journal of Autism and Developmental Disorders, 42*(9), 1870–1883.

Perner, J. (1996). Simulation as explicitation of predication-implicit knowledge about the mind: Arguments for a simulation-theory mix. In *Theories of theories of mind* (pp. 90–104). Cambridge: Cambridge University Press.

Pfundmair, M., Rimpel, A., Duffy, K., & Zwarg, C. (2018). Oxytocin blurs the self-other distinction implicitly but not explicitly. *Hormones and Behavior, 98*, 115–120. https://doi.org/10.1016/j.yhbeh.2017.12.016

Preston, S. D., & De Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences, 25*(1), 1–20.

Reniers, R. L., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A questionnaire of cognitive and affective empathy. *Journal of Personality Assessment, 93*(1), 84–95.

Richer, J. (2001). The insufficient integration of self and other in autism. In J. M. Richer & S. Coates (Eds.), *Autism: The search for coherence in London*. Vancouver, BC: University of British Columbia.

Riecansky, I., Lengersdorff, L., Pfabigan, D., & Lamm, C. (2020). Increasing self-other bodily overlap increases sensorimotor resonance to others' pain. *Cognitive, Affective, & Behavioral Neuroscience, 20*, 19.

Rochat, P., & Striano, T. (2000). Perceived self in infancy. *Infant Behavior and Development, 23*(3-4), 513–530.

Rogers, K., Dziobek, I., Hassenstab, J., Wolf, O. T., & Convit, A. (2007). Who cares? Revisiting empathy in Asperger syndrome. *Journal of Autism and Developmental Disorders, 37*(4), 709–715.

Rogers, S. J. (1999). An examination of the imitation deficit in autism. In J. Nadel & G. Butterworth (Eds.), *Cambridge studies in cognitive perceptual development. Imitation in infancy* (pp. 254–283). Cambridge: Cambridge University Press.

Rogers, S. J., Bennetto, L., McEvoy, R., & Pennington, B. F. (1996). Imitation and pantomime in high-functioning adolescents with autism spectrum disorders. *Child Development, 67*(5), 2060–2073.

Rogers, S. J., & Pennington, B. F. (1991). A theoretical approach to the deficits in infantile autism. *Development and Psychopathology, 3*(2), 137–162.

Ruissen, M. I., & de Bruijn, E. R. (2015). Is it me or is it you? Behavioral and electrophysiological effects of oxytocin administration on self-other integration during joint task performance. *Cortex, 70*, 146–154.

Ruiz-Mirazo, K., Peretó, J., & Moreno, A. (2004). A universal definition of life: Autonomy and open-ended evolution. *Origins of Life and Evolution of the Biosphere, 34*(3), 323–346. https://doi.org/10.1023/B:ORIG.0000016440.53346.dc

Sagi, A., & Hoffman, M. L. (1976). Empathic distress in the newborn. *Developmental Psychology, 12*(2), 175.

Santiesteban, I., Banissy, M. J., Catmur, C., & Bird, G. (2012). Enhancing social ability by stimulating right temporoparietal junction. *Current Biology, 22*(23), 2274–2277. https://doi.org/10.1016/j.cub.2012.10.018

Santiesteban, I., White, S., Cook, J., Gilbert, S. J., Heyes, C., & Bird, G. (2012). Training social cognition: From imitation to theory of mind. *Cognition, 122*(2), 228–235.

Shamay-Tsoory, S. G. (2011). The neural bases for empathy. *The Neuroscientist, 17*(1), 18–24.

Shamay-Tsoory, S. G., & Abu-Akel, A. (2016). The social salience hypothesis of oxytocin. *Biological Psychiatry, 79*(3), 194–202.

Silani, G., Lamm, C., Ruff, C. C., & Singer, T. (2013). Right supramarginal gyrus is crucial to overcome emotional egocentricity bias in social judgments. *The Journal of Neuroscience, 33*(39), 15466–15476. https://doi.org/10.1523/jneurosci.1488-13.2013

Singer, T., Seymour, B., O'doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science, 303*(5661), 1157–1162.

Smith, A. (2006). Cognitive empathy and emotional empathy in human behavior and evolution. *The Psychological Record, 56*(1), 3–21.

Smith, I. M., & Bryson, S. E. (1994). Imitation and action in autism: A critical review. *Psychological Bulletin, 116*(2), 259.

Spengler, S., Bird, G., & Brass, M. (2010). Hyperimitation of actions is related to reduced understanding of others' minds in autism spectrum conditions. *Biological Psychiatry, 68*(12), 1148–1155.

Spiker, D., & Ricks, M. (1984). Visual self-recognition in autistic children: Developmental relationships. *Child Development, 55*, 214–225.

Steinbeis, N. (2016). The role of self–other distinction in understanding others' mental and emotional states: Neurocognitive mechanisms in children and adults. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 371*(1686), 20150074. https://doi.org/10.1098/rstb.2015.0074

Sui, J., Chechlacz, M., Rotshtein, P., & Humphreys, G. W. (2013). Lesion-symptom mapping of self-prioritization in explicit face categorization: Distinguishing hypo-and hyper-self-biases. *Cerebral Cortex, 25*(2), 374–383.

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin, 121*(3), 371.

Taylor, S. E. (2006). Tend and befriend: Biobehavioral bases of affiliation under stress. *Current Directions in Psychological Science, 15*(6), 273–277.

Thirioux, B., Wehrmann, M., Langbour, N., Jaafari, N., & Berthoz, A. (2016). Identifying oneself with the face of someone else impairs the egocentered visuo-spatial mechanisms: A new double mirror paradigm to study self–other distinction and interaction. *Frontiers in Psychology, 7*, 1283. Retrieved from https://www.frontiersin.org/article/10.3389/fpsyg.2016.01283

Tomova, L., von Dawans, B., Heinrichs, M., Silani, G., & Lamm, C. (2014). Is stress affecting our ability to tune into others? Evidence for gender differences in the effects of stress on self-other distinction. *Psychoneuroendocrinology, 43*, 95–104. https://doi.org/10.1016/j.psyneuen.2014.02.006

Tong, F., & Nakayama, K. (1999). Robust representations for faces: Evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance, 25*(4), 1016.

Uddin, L. Q., Davies, M. S., Scott, A. A., Zaidel, E., Bookheimer, S. Y., Iacoboni, M., & Dapretto, M. (2008). Neural basis of self and other representation in autism: An FMRI study of self-face recognition. *PLoS One, 3*(10), e3526.

Uzefovsky, F., & Baron-Cohen, S. (2018). Role taking. In M. H. Bornstein (Ed.), *The SAGE encyclopedia of lifespan human development*. Thousand Oaks, CA: SAGE Publications.

Uzefovsky, F., Bethlehem, R. A., Shamay-Tsoory, S., Ruigrok, A., Holt, R., Spencer, M., … Bullmore, E. (2019). The oxytocin receptor gene predicts brain activity during an emotion recognition task in autism. *Molecular Autism, 10*(1), 12.

Uzefovsky, F., & Knafo-Noam, A. (2017). Empathy development throughout the life span. In *Social cognition: Development across the life span* (pp. 71–97). New York, NY: Routledge.

von Dawans, B., Kirschbaum, C., & Heinrichs, M. (2011). The Trier Social Stress Test for Groups (TSST-G): A new research tool for controlled simultaneous social stress exposure in a group format. *Psychoneuroendocrinology, 36*(4), 514–522. https://doi.org/10.1016/j.psyneuen.2010.08.004

Wang, Y., & Hamilton, A. F. C. (2012). Social top-down response modulation (STORM): A model of the control of mimicry in social interaction. *Frontiers in Human Neuroscience, 6*, 153.

Wellman, H. (1990). *The child's theory of mind. A Bradford Book*. Cambridge, MA: MIT Press.

Whiten, A., & Brown, J. (1998). Imitation and the reading of other minds: Perspectives from the study of autism, normal children and non-human primates. In *Intersubjective communication and emotion in early ontogeny* (pp. 260–280). Cambridge: Cambridge University Press.

Williams, D., & Happé, F. (2009a). Pre-conceptual aspects of self-awareness in autism spectrum disorder: The case of action-monitoring. *Journal of Autism and Developmental Disorders, 39*(2), 251–259.

Williams, D. M., & Happé, F. (2009b). What did I say? Versus what did I think? Attributing false beliefs to self amongst children with and without autism. *Journal of Autism and Developmental Disorders, 39*(6), 865–873.

Williams, J. H., Whiten, A., Suddendorf, T., & Perrett, D. I. (2001). Imitation, mirror neurons and autism. *Neuroscience & Biobehavioral Reviews, 25*(4), 287–295.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods, 8*(8), 665.

Yirmiya, N., Erel, O., Shaked, M., & Solomonica-Levi, D. (1998). Meta-analyses comparing theory of mind abilities of individuals with autism, individuals with mental retardation, and normally developing individuals. *Psychological Bulletin, 124*(3), 283.

Young, S. K., Fox, N. A., & Zahn-Waxler, C. (1999). The relations between temperament and empathy in 2-year-olds. *Developmental Psychology, 35*(5), 1189.

Zhao, W., Yao, S., Li, Q., Geng, Y., Ma, X., Luo, L., … Kendrick, K. M. (2016). Oxytocin blurs the self-other distinction during trait judgments and reduces medial prefrontal cortex responses. *Human Brain Mapping, 37*(7), 2512–2527.

Zwaigenbaum, L., Bryson, S., Rogers, T., Roberts, W., Brian, J., & Szatmari, P. (2005). Behavioral manifestations of autism in the first year of life. *International Journal of Developmental Neuroscience, 23*(2-3), 143–152.

The Evolution of Mentalizing in Humans and Other Primates

Christopher Krupenye

Human hyper-sociality is remarkable when considered in comparative context (Boyd, 2006; Sterelny, 2019; Tomasello, Melis, Tennie, Wyman, & Herrmann, 2012). We cooperate and compete on unprecedented scales and in unique and flexible ways, we exhibit systems of communication not seen in any other species, and we are deeply cultural. On a proximate level, many of the unique features that define human sociality are made possible by our theory of mind (Adolphs, 2009; Banaji & Gelman, 2012; Tomasello, Carpenter, Call, Behne, & Moll, 2005; Tremblay, Sharika, & Platt, 2017; Kovacs et al, 2010). Theory of mind (or mentalizing, mental state attribution, mind-reading) is the ability to ascribe mental states, such as desires and beliefs, to others (Premack & Woodruff, 1978). We consider others' motives and perspectives when we communicate or teach, or when we attempt to cooperate or to deceive. We even delineate cultural group membership on the basis of shared versus differing beliefs. A prominent hypothesis has long been that the absence of theory of mind (or particular features of theory of mind) in other species may explain the apparent gulf in social complexity between humans and our closest primate relatives (e.g., Herrmann, Call, Hernandez-Lloreda, Hare, & Tomasello, 2007). Given the centrality of theory of mind to human social life, for more than 40 years, researchers have endeavored to clarify its evolutionary origins and to determine whether this capacity is indeed unique to our species (Krupenye & Call, 2019). The term "theory of mind" was even defined by Premack and Woodruff (1978) in their quest to determine whether our closest phylogenetic relative, the chimpanzee, possesses one.

As Tinbergen (1963) adumbrated beautifully, a comprehensive understanding of the origins of a behavior requires its investigation from four inter-linked perspectives. On a proximate level, we must elucidate the causal, in this case cognitive (or even more basally, neural), mechanisms that underlie the phenomenon of interest,

C. Krupenye (✉)
Department of Psychological & Brain Sciences, Johns Hopkins University,
Baltimore, MD, USA

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_6

as well as the ontogenetic processes that drive its development. On an ultimate (evolutionary) level, we should be concerned with the distribution of these causal mechanisms across phylogeny as well as the behavior's function or adaptive significance—the ways in which the behavior improves the survival and reproductive success of its bearers and was thus favored by natural selection. In the present paper, I will explore the evolutionary origins of human theory of mind. This exploration will focus predominantly on the distribution of key constituent mechanisms and evolutionary precursors across primate phylogeny and then briefly on the selective pressures that likely shaped the social minds of humans and our closest relatives. The first approach will allow us to reconstruct the evolutionary history of theory of mind and the second to determine how it came to be.

Charting the Evolutionary History of Human Theory of Mind

Cognitive traits do not fossilize. Phylogenetic comparisons of living species are therefore essential not only for identifying potentially unique features of human theory of mind but also precursors that are shared across closely related taxa and were likely present in their common ancestor. Through this approach, we can reconstruct the cognitive phenotype of our common ancestors at different points in primate evolutionary history, and chart patterns of change (MacLean et al., 2012). Which features of human theory of mind are shared across primates, suggesting particularly deep phylogenetic roots? Which capacities can be found only in species most closely related to humans (e.g., monkeys and apes or only in apes), having been built by evolution on those earliest primate-wide foundations? Finally, which features evolved in the human lineage alone, over the last 6–9 million years since its divergence from the other apes? Phylogenetic reconstruction allows us to simulate the sequential process by which evolution assembled our cognition and, in doing so, can provide useful insights into the nature and architecture of its underlying mechanisms.

Evolutionary forces work by building on or modifying existing structures (Darwin, 1859). Populations exhibit natural variation in traits, and variants that permit individuals to better survive and reproduce will be favored, becoming more prolific within the population. Over generations, consistent selective pressures can slowly drive the enhancement of a trait (e.g., increases in or elaboration of a cognitive ability). Accordingly, ancestral forms are often expected to differ from their modern descendants more in degree than in kind, reflecting the largely gradual nature of evolutionary processes. Evolutionary accounts of cognition therefore favor (in many but of course not all cases) cognitive mechanisms that vary continuously across species. In the case of theory of mind, it may be that social cognition reflects a continuum of computations about others' perspectives or mental states (e.g., what others perceive vs. know vs. believe) that increased in complexity throughout our evolutionary history. In the following sections, I will attempt to reconstruct the cognitive abilities of our ancestors at key points in primate evolutionary history, and consider the potential continuity of these mechanisms.

What Is Common to All Primates?

The most basal divergence in primate evolutionary history occurred around 77 million years ago (mya) when the common ancestors of all living primates split into two lineages: the *strepsirrhines*, which eventually produced lemurs, galagos, and lorises, and the *haplorhines*, which eventually produced monkeys and apes (including humans; Fig. 1) (Steiper & Young, 2006). Although less experimental research effort has been devoted to understanding the social cognitive abilities of strepsirrhines as compared with haplorhines, the limited existing work suggests that a very basic sensitivity to others may be common to all primates.

Across most taxa studied, primates are responsive to others' gaze (Rosati & Hare, 2009). Both strepsirrhines and haplorhines follow the gaze of their conspecifics, an act that could facilitate detection of food, predators, competitors, and mates—and which is fundamental to determining others' visual perspectives (Ruiz, Gomez, Roeder, & Byrne, 2009; Shepherd & Platt, 2008; Tomasello, Call, & Hare,

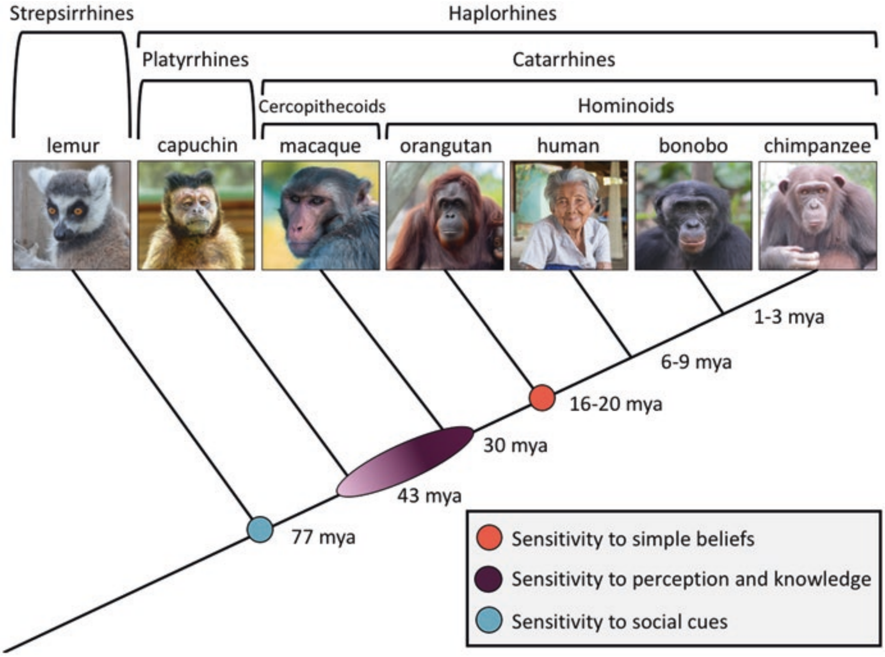


Fig. 1 Phylogeny of primate theory of mind. Comparative research suggests that the roots of human mentalizing extend deep into primate phylogeny

1998). In competitive contexts, several species of lemur (as well as monkeys and apes) also preferentially steal food that a competitor is oriented away from over food that is in the competitor's plain view (Brauer, Call, & Tomasello, 2007; Bray, Krupenye, & Hare, 2014; Canteloup, Piraux, Poulin, & Meunier, 2016; Flombaum & Santos, 2005; Hare, Call, Agnetta, & Tomasello, 2000; Hare, Call, & Tomasello, 2001, 2006; MacLean, Sandel, Reddy, et al., 2013; Sandel, MacLean, & Hare, 2011).

However, another set of experiments by Bray et al. (2014) suggests that strepsirrhine social cognition may be much more limited than that of other primates. Despite the fact that ring-tailed lemurs (*Lemur catta*) performed well in other competitive tasks (e.g., MacLean, Sandel, Reddy, et al., 2013; Sandel et al., 2011), Bray et al. (2014) found little evidence that they could model others' perspective by integrating information about what others could hear with information about what they could see. Although lemurs were less likely to approach a human competitor facing them as compared with one facing away, when the competitor *was* facing away, they did not preferentially steal from a silent location over a noisy one (Bray et al., 2014). Monkeys and apes, in contrast, selectively avoid noisy options when a competitor cannot see them (Melis, Call, & Tomasello, 2006; Santos, Nissen, & Ferrugia, 2006).

Taken together, these findings suggest that lemurs—and perhaps the common ancestors of all living primates—are endowed with mechanisms for detecting and responding to coarse metrics of visual orientation but that they may not be capable of integrating information across modalities or computing, in any greater detail, the perspectives of others.

What Is Unique to Monkeys and Apes?

Monkeys and apes, on the other hand, have shown much greater sensitivity to human eye orientation and an ability, in competitive tasks, to simultaneously exploit information about visual and auditory perspective (Flombaum & Santos, 2005; Melis et al., 2006; Santos et al., 2006; Tomasello, Hare, Lehmann, & Call, 2007; but see Brauer, Call, & Tomasello, 2008). These findings suggest that the haplorhine lineage has experienced an elaboration of primate social cognition since its divergence from the strepsirrhines. Other studies have helped to clarify the nature and degree of this elaboration.

Among haplorhines, there have been two major divergences (Fig. 1). Roughly 43 mya, the platyrrhines (new world monkeys) split from the catarrhines (old world monkeys and apes), and later, around 30 mya, the catarrhines further splintered into cercopithecoids (old world monkeys) and hominoids (apes, including humans) (Steiper & Young, 2006). Across the representative platyrrhines, cercopithecoids, and hominoids that have been tested, there is evidence that haplorhines, broadly, treat agents as goal-directed. For example, they detect certain cues of animacy and agency, discriminate similar movements that are underlain by different intentions, complete others' failed actions, and anticipate the outcome of others' goal-directed behavior (Anderson, Kuroshima, Takimoto, & Fujita, 2013; J. Burkart, Kupferberg, Glasauer, & van Schaik, 2012; Buttelmann, Carpenter, Call, & Tomasello, 2007; Buttelmann, Schütte, Carpenter, Call, & Tomasello, 2012; Call, Hare, Carpenter, & Tomasello, 2004; Call & Tomasello, 1998; Canteloup & Meunier, 2017; Drayton & Santos, 2014; Kano & Call, 2014; Krupenye & Hare, 2018; Krupenye, Tan, & Hare, 2018; Kupferberg, Glasauer, & Burkart, 2013; Myowa-Yamakoshi & Matsuzawa, 2000; Phillips, Barnes, Mahajan, Yamaguchi, & Santos, 2009; Rochat, Serra, Fadiga, & Gallese, 2008; Tomasello & Carpenter, 2005; Warneken, Hare, Melis, Hanus, & Tomasello, 2007; Warneken & Tomasello, 2006; Yamamoto, Humle, & Tanaka, 2012). This body of work suggests that haplorhines share a basic capacity to identify the goals and intentions that motivate others' actions, although it remains unclear whether they do so by tracking behavior or mental states. It is also unclear whether this capacity is unique to haplorhines, since complementary work has not been completed with strepsirrhine primates.

Haplorhines have also demonstrated a more complex and integrated understanding of others' perspectives than strepsirrhines, although data are largely restricted to old world monkeys (especially rhesus macaques, *Macaca mulatta*) and apes (especially chimpanzees, *Pan troglodytes*) (except, e.g., Burkart & Heschl, 2007; Hare, Addessi, Call, Tomasello, & Visalberghi, 2003; Meunier, 2017). As described above, rhesus macaques and chimpanzees are both able to integrate information about what a competitor can perceive from multiple modalities (specifically, seeing and hearing) to successfully pilfer food (Melis et al., 2006; Santos et al., 2006). They also show a richer understanding of others' gaze. All great apes, and in some cases rhesus macaques, follow gaze geometrically, around barriers, and check back with an actor when they cannot identify the target of her gaze (Bettle & Rosati, 2019; Brauer, Call, & Tomasello, 2005; Horton & Caldwell, 2006; Okamoto-Barth, Call, & Tomasello, 2007; Povinelli & Eddy, 1996). Chimpanzees and rhesus macaques are also able to track, in some sense, whether someone can see something, preferentially pilfering food that a competitor cannot see (Brauer et al., 2007; Crockford, Wittig, Mundry, & Zuberbuhler, 2012; Flombaum & Santos, 2005; Hare et al., 2000, 2006; Karg, Schmelz, Call, & Tomasello, 2015a, 2015b; Melis et al., 2006).

Apes and rhesus macaques can even track, in some sense, whether an actor has *previously* seen something and is therefore aware of its location. For example, chimpanzees preferentially steal food that a competitor has not seen being hidden (Hare et al., 2001; Kaminski, Call, & Tomasello, 2008). Chimpanzees and macaques also expect an actor to retrieve a desirable object from a concealed location if the actor previously witnessed the object being stashed there (Drayton & Santos, 2018; Kaminski et al., 2008; Marticorena, Ruiz, Mukerji, Goddu, & Santos, 2011). They also infer the target of an agent's attention based on what the agent has previously seen: in one particularly elegant paradigm (Tomasello & Haberl, 2003), chimpanzees, bonobos, and rhesus macaques watched as an object was placed beside a human actor either in or out of view of the actor (Drayton & Santos, 2017; MacLean & Hare, 2012). The actor then turned toward the object and expressed surprise. Fascinatingly, subjects of all three species responded differently to the knowledgeable versus ignorant actors: they searched for an alternative target of the actor's gaze more often or more quickly when they knew that she was already familiar with the object in front of her. These various results indicate that (at least some) catarrhine primates are able to closely track an actor's attention and his awareness of objects, including those that are no longer visible or that he only encountered in the past. This skill, in catarrhines, appears to be at least functionally equivalent to representing the actor as "seeing" and "knowing" the object's location (but see below for further discussion of mechanisms).

Critically, however, several studies suggest that rhesus macaques may struggle to track others' perspectives if they conflict with the monkey's own (e.g., Lorincz et al., 2005; Marticorena et al., 2011; Martin & Santos, 2014). In one such study, designed to assess macaques' expectations of actors with true versus false beliefs (based on Onishi & Bailargeon, 2005), Marticorena et al. (2011) presented macaques with a display in which a human actor watched an object move from one box to another. When the actor witnessed this transfer (and had the same true belief about the object's location as the monkeys), monkeys expected her to reach into the box containing the object: they looked longer (indicative of surprise or an unexpected outcome) when she instead reached into the empty box. In a second study, the actor witnessed the object entering one of the boxes but then could not see as the object moved from the first box to the second. Here, the monkeys held a true belief that the object was in the second box while the actor held a conflicting false belief that it remained in the first. If the monkeys could represent the actor's false belief, they should expect her to reach into the (now empty) first box. However, they instead seemed to have no prediction about where the actor would search: the duration of their looking did not differ significantly if she reached into the first box or the second.

Perhaps most strikingly, in a related study, Horschler, Santos, and MacLean (2019) demonstrated that the same monkeys show no clear prediction, even when the monkey and actor share a true belief about an object's location, *if* the actor has missed (irrelevant) intervening information about the object's movement history. For example, if the actor witnessed an object enter a box but could not see as the object briefly exited and returned to that same box, monkeys showed comparable looking whether the actor then reached into the empty box or the one containing the object.

This body of data suggest that rhesus macaques are able to track, in either behavioral (Penn & Povinelli, 2007; Povinelli & Vonk, 2003) or mentalistic terms (Marticorena et al., 2011), what others can perceive and also what others have perceived previously and therefore, in some sense, know. Interestingly, monkeys appear to have signature failures in predicting an agent's actions as soon as the agent's perspectives differ from the monkey's own. What cognitive mechanisms might explain this body of findings? There are two particularly prominent hypotheses. Martin and Santos (2016) recently argued that monkeys may represent awareness relations, allowing them to track the aspects of reality that others have detected—but nothing more. This hypothesis postulates that monkeys establish an awareness relation between an agent and an object when they observe the agent encountering the object in a particular location (e.g., seeing or hearing it). Awareness relations persist even after the object becomes imperceptible (e.g., monkeys encode an agent as being aware of things the agent has seen hidden). However, if ever the object moves while the agent is not attending, including irrelevant movements as in the Horschler et al. (2019) study, the awareness relation is broken or turned off. The authors argue that monkeys then have no representation whatsoever about the relation between the agent and the object and therefore also have no expectation that the agent will pursue, find, or interact with the object in any capacity. This is why monkeys showed patterns of looking indicative of no expected search in either the false belief condition of Marticorena et al. (2011), where the actor missed the object moving from one box to the other, or the Horschler et al. (2019) condition in which the actor simply missed the object briefly leaving and then returning to its original location.

The alternative account posits that monkeys can represent states along the lines of knowledge and ignorance but not beliefs. Understanding others as having beliefs, as humans do, requires simultaneously representing two views of the world—one's own, which aligns with reality, and the conflicting view of another (Krupenye & Call, 2019). In contrast, both the awareness relations and knowledge-ignorance accounts argue that monkeys are able to represent just one view of the world—their own—but that they can keep track of the aspects of it that others also have access to. Thus, similar to the awareness relations hypothesis, the knowledge-ignorance account suggests that monkeys are able to track the bits of reality that others *are* aware of, such as what others can see or have seen. The accounts differ fundamentally, though, in that the knowledge-ignorance account argues that monkeys can additionally track the bits of reality that others are *not* aware of—that they can represent that an agent cannot see something or does not know about it. According to this account, monkeys show no expectations about where the actor will search in the false belief condition of Marticorena et al. (2011) because they recognize that the actor has not witnessed the object's movement and attribute to that actor a state of ignorance (and, of course, ignorant actors should be expected to search at random or not at all; Baillargeon, Scott, & He, 2010). In the Horschler et al. (2019) study, monkeys initially encode the actor as being aware that the object has been hidden in a particular location. They also recognize that the actor is unable to witness the object's subsequent removal and re-hiding in that same location, and attribute to the actor a state of ignorance about these events. According to the knowledge-ignorance account, they have no prediction about the actor's search behavior in this condition because of an issue with integration: they struggle to integrate their original representation (actor is knowledgeable that the object is in location 1) with their subsequent representation (actor is ignorant that the object is in location 1) and they act on the basis of this later representation. At this point, both the awareness relations and knowledge-ignorance hypotheses are able to account for the full body of data in monkeys, and future work will be necessary to distinguish them.

Returning to our phylogenetic reconstruction (Fig. 1), findings from monkeys and apes suggest that the capacity to treat others' actions as goal-directed likely extends back to at least the haplorrhine common ancestor (i.e., the common ancestor of all monkeys and apes). It is important to note that a paucity of data on strepsirrhines currently obscures our ability to determine whether this skill was present even earlier. The capacity to track, in some sense, what others can see and hear and what they know on the basis of seeing and hearing may date back to the haplorrhine common ancestor as well: there is some evidence of these skills in our most distant monkey relatives, the platyrrhines (e.g., Burkart & Heschl, 2007; Defolie, Malassis, Serre, & Meunier, 2015; Hare et al., 2003). However, the strongest evidence comes from closer relatives, macaques and apes—suggesting that these capacities were likely present at least in the subsequent catarrhine ancestor of old world monkeys and apes. As described above, additional experiments are necessary to clarify the particular cognitive representations that underlie these skills in rhesus macaques and likely our common ancestor with them.

What Is Unique to Apes?

What has changed since that time? Thirty million years ago, the hominoid (ape) lineage diverged from cercopithecoids (old world monkeys, like macaques). Extant apes consist of the lesser apes—the gibbons and siamangs of the family Hylobatidae whose social cognition we know little about (Horton & Caldwell, 2006; Liebal & Kaminski, 2012)—and the great apes of the family Hominidae, of which our species is a member (Fig. 1). Among the great apes, our more distant relatives are the orangutans (genus *Pongo*, common ancestor 16–20 mya) followed by the gorillas (genus *Gorilla*, common ancestor 7–9 mya) (Steiper & Young, 2006). Our very closest relatives are chimpanzees and bonobos, who are equally closely related to us and with whom we share a 6–9 million year old common ancestor (Muller, Wrangham, & Pilbeam, 2017). As our closest relatives, great apes—and especially chimpanzees and bonobos—therefore provide the deepest available insights into the features of human psychology that preceded the origin of our lineage (Krupenye, MacLean, & Hare, 2017). Excitingly, recent discoveries across the great apes raise the possibility that their social cognitive abilities exceed those of rhesus macaques (and our catarrhine common ancestor), and more closely mirror our own social minds in fundamental ways.

It has long been believed that apes, like rhesus macaques, are unable to represent beliefs or perspectives that differ from their own (Call & Tomasello, 1999, 2008). However, recent evidence challenges this view, demonstrating that apes can predict—and also respond appropriately to—the behavior of an actor who holds a false belief. For example, Krupenye, Kano, Hirata, Call, and Tomasello (2016) presented captive chimpanzees, bonobos, and orangutans with videos of classic false belief manipulations embedded within dramatic social conflicts of the sort that pervade ape social life (Kano, Krupenye, Hirata, & Call, 2017; Southgate, Senju, & Csibra, 2007). In one study, for example, a human actor was searching for a gorilla-like character who had hidden in one of two haystacks. In one of the critical false belief manipulations, the actor witnessed the gorilla hiding in one location and then briefly left the scene. While away, the gorilla moved to the other location before ultimately fleeing the scene himself. An eye-tracker noninvasively recorded apes' gaze and revealed that, when the actor subsequently returned, apes looked to the location where the actor had last seen the gorilla, in anticipation of his search: that is, they predicted that he would search in accordance with his false belief, even though the apes themselves knew that the gorilla was no longer there. The original study and a subsequent control ruled out a variety of low-level perceptual or domain-general explanations (Kano, Krupenye, Hirata, Call, & Tomasello, 2017; Krupenye, Kano, Hirata, Call, & Tomasello, 2017).

Excitingly, an additional experience-projection eye-tracking study has since replicated and extended these findings, and provided critical evidence to bolster the view that apes' successful anticipation reflects attribution of mental states rather than sophisticated reading of behavior (Kano, Krupenye, Hirata, Tomonaga, & Call, 2019). Apes and an actor witnessed an object being hidden in one location before the actor scuttled behind a novel barrier and the object was subsequently moved and removed. Apes anticipated that the actor would seek the object in its original location (accordant with attribution of a false belief) if the apes had previously personally experienced the same barrier as opaque but not if they had experienced the barrier as translucent. These results are consistent with the possibility that apes leveraged their own perceptual experience to attribute differential perceptual access (and potentially beliefs) (but see Lurz, 2009), and correctly anticipate the actor's subsequent actions (see also Heyes, 1998; Karg et al., 2015b; Lurz, Krachun, Mahovetz, Wilson, & Hopkins, 2018; Meltzoff & Brooks, 2008; Penn & Povinelli, 2007; Senju, Southgate, Snape, Leonard, & Csibra, 2011; Whiten, 2013). Similar manipulations should be pursued with monkeys to strengthen the case that they share this apparently mentalistic understanding of gaze.

Apes' sensitivity to others' beliefs can also influence their actions as Buttelmann, Buttelmann, Carpenter, Call, and Tomasello (2017) have shown. In their paradigm, apes provided different help to an actor attempting to open an empty box, depending on whether or not the actor believed that an object was still in the box. Interestingly, although the cohort of tasks that preceded the eye-tracking and helping studies have largely been viewed as providing clear evidence that apes cannot represent others' beliefs, their results are actually much more mixed. As Horschler et al. (2019) point out, the findings of several studies, in which apes search for food in one of two containers, can be explained by cue-following: apes continue to follow cooperative, competitive, or color cues that are indicative of the food's location in true belief conditions, even when the signaler has a false belief about that location (Call & Tomasello, 1999; Krachun, Carpenter, Call, & Tomasello, 2009, 2010). Interestingly, in one of these studies (but not another), even when apes reliably followed this cue, they were more likely to glance at the alternative option in the false belief condition than the true belief condition—potentially indicative of some implicit understanding (Krachun et al., 2009). Hare et al. (2001) also found that subordinate chimpanzees were more likely to acquire a piece of food when competing against a dominant who had a false belief about its location as compared with a true belief. The likelihood that chimpanzees even entered the competition arena was also marginally greater if their competitor was misinformed rather than knowledgeable (but this trend was not significant). However, these behaviors were mirrored in conditions in which the competitor was ignorant (as compared to knowledgeable). Thus, although appeals to false belief attribution are not necessary to explain these results, the results are also not inconsistent with such a mechanism.

Similar findings were reported by Kaminski et al. (2008). These authors confronted chimpanzees with a turn-taking food competition task in which competitors could seek a single high-value food item hidden among several cups, or opt for a safe lower-value option. The subject was always knowledgeable and also allowed to track the competitor's awareness, which was manipulated across conditions. When the competitor chose first, chimpanzees were more likely to pursue the high-value option if they knew that their competitor was misinformed (and likely wouldn't have found the food) than when they knew that their competitor was knowledgeable. However, they were also more likely to pursue the high-value option if their competitor had seen it hidden in one location but had not subsequently seen it removed and replaced in that same location, than if the competitor had witnessed all of these actions—a condition children found challenging as well. Their performance on this control could be explained, however, if apes struggled to integrate conflicting representations of the actor's belief about a single object in a single location—and acted on the basis of the more recent attribution of ignorance (as monkeys may have done in Horschler et al., 2019).

Together, this body of work indicates that, at least under minimally demanding conditions, apes are able to predict (and respond appropriately to) the actions of an agent who is mistaken about the location of a hidden object. These findings suggest that ape social cognition is undergirded by representations more sophisticated than awareness relations. The awareness relations hypothesis predicts that apes should reliably fail all tasks in which an agent's view of the world conflicts with the subject's own (and with reality). However, findings from recent false belief tasks imply instead that apes' representations of others' beliefs persist, even when they are no longer congruent with reality (the time when awareness relations are proposed to turn off). These and other data from apes are therefore best explained by the hypothesis that apes can represent, in some sense, an actor's knowledge (i.e., that an actor has seen something hidden and will search for it there), ignorance (i.e., that an actor has not seen something hidden and is unlikely to search for it there), and simple forms of false beliefs about an object's location (i.e., that an actor has seen something hidden and will search for it there, even though the subject knows that it has since been moved). That these representations are mentalistic is bolstered by apes' success on experience-projection tasks about seeing and believing (Heyes, 1998; Kano et al., 2019; Karg et al., 2015b; Penn & Povinelli, 2007; Whiten, 2013). This body of work suggests that apes' social cognition is non-egocentric and that, in some respect, apes can simultaneously represent two views of the world—their own, and the conflicting view of another (Krupenye & Call, 2019).

That being said, there remain several plausible accounts of the mechanisms that underlie this rich social understanding. These accounts are of at least two particularly prominent varieties. The first is that nonhuman apes, like humans, can represent beliefs and other propositional attitudes as such, in their full sense. This account predicts that apes will succeed on the full range of first-order false belief tasks that exist beyond the change-of-location paradigms in which they've so far been tested.

The primary alternative is that they may have what Butterfill and Apperly (2013) have termed a minimal theory of mind. According to this account, apes are able to detect when an agent encounters an object (i.e., when the object is in the agent's perceptual field) and additionally they can represent that the agent will register the object in the last place she encountered it. These registrations, or belief-like states, are akin to beliefs about an object's location; representing them therefore allows apes to correctly predict behavior in a wide range of cases when an actor has a true or false belief related to the location of an object. However, minimal theory of mind suffers a signature limitation: belief-like states encode location information that is specific to objects familiar to both the agent and the minimal mindreader; they therefore cannot account for false beliefs about the identity or aspect of an object (e.g., that some agents may falsely believe that Clark Kent and Superman are two different individuals). Accordingly, and in contrast to the propositional attitude hypothesis, this account predicts that apes will fail change-of-identity false belief tasks—a key future direction for this line of research.

Taken together, this body of work suggests that great apes are able to track, in some sense, at least simple false beliefs about the location of hidden objects. Such a capacity has not been documented in monkeys (or any other species), despite several experimental attempts involving minimally demanding gaze-based paradigms (Marticorena et al., 2011; Martin & Santos, 2014). Together, these findings raise the possibility that the ability to represent others' perspectives, even when they conflict with one's own, evolved uniquely in the ape lineage, at least 16–20 million years ago.

In order to confirm this hypothesis, several follow-up experiments with monkeys should be pursued. Although the false belief tests that monkeys have failed are not dissimilar to those that apes have passed, there are a few key distinctions. First, gaze-based studies with apes have relied on highly engaging social stimuli specifically designed (1) to maximize apes' interest and the ecological validity of these paradigms and (2) to ensure that apes are highly motivated to track and predict the behavior of the agents involved (Kano et al., 2019; Kano, Krupenye, Hirata, & Call, 2017; Krupenye et al., 2016). However, attention may not be responsible for divergent performance between monkeys and apes, given that monkeys correctly predicted behavior in true belief conditions not unlike the false belief conditions they failed (e.g., Marticorena et al., 2011). Second, these same ape studies are based on anticipatory looking—active prediction in advance of the actor's search—as opposed to violation of expectation—a reaction to the actor's search. It's possible that, as a proactive rather than reactive measure, anticipatory looking is more sensitive to attribution of beliefs. Finally, a potentially key element of the Krupenye et al. (2016) and Kano et al. (2019) studies is that the target object was ultimately removed from the scene. This design decision was made to ensure that participants would not exhibit a reality bias, simply looking to the location where the object was hidden as opposed to the location the actor believed that the object was hidden. It's possible that monkeys did not show differential looking in the false belief condition of Marticorena et al. (2011) because they found it somewhat expected that an actor would search incorrectly, where she falsely believed an object to be, *or* that the actor would search correctly, where the monkey knew that the object was. The monkey's own knowledge may have muddied their predictions or created interference that impacted their looking times. Thus, an important test of whether monkeys do indeed differ from apes in this fundamental respect will be examining whether they can succeed in predicting an actor's search on a false belief condition in which the object has been removed before that search. Finding that monkeys fail both anticipatory looking tasks with dynamic stimuli and violation-of-expectation paradigms where the target object has been removed would greatly strengthen the claim that apes have uniquely evolved the capacity to understand others' perspectives, even when they differ from one's own.

What Is Unique to Humans?

Comparative data paint a picture of a 6–9 million year old common ancestor of chimpanzees, bonobos, and humans that possessed a political mind and a rich understanding of its dynamic social world. Our ancestors closely tracked third-party interactions and made strategic social decisions on the basis of this information (e.g., Krupenye & Hare, 2018; Wittig, Crockford, Langergraber, & Zuberbuhler, 2014), and they represented others' perspectives in a relatively rich sense (e.g., Kano et al., 2019; Krupenye et al., 2016). And yet, much has subsequently changed to produce the unrivaled sophistication of our own social minds and the unrivaled social and cultural complexity that uniquely characterizes our species. There are several notable cognitive candidates that may explain, on a proximate level, the apparent gulf between humans and our closest living relatives.

First, as described above, it remains unknown whether apes can represent propositional attitudes as such or whether they are limited to attributing belief-like states. Second, we don't know whether apes, like humans, are capable of level II perspective-taking—that is, understanding or imagining *how* something looks from another's perspective (Flavel, Everett, Croft, & Flavel, 1981; Karg, Schmelz, Call, & Tomasello, 2016; Moll & Meltzoff, 2011). Third, we don't know the extent to which apes can engage in recursive mind-reading (Corballis, 2011). Attributing basic belief states is a case of first-order theory of mind ("I know that you believe …"). Humans, however, are capable of representing much higher levels of embedding of mental states—perhaps at least seven (O'Grady, Kliesch, Smith, & Scott-Phillips, 2015). It is possible that by providing an efficient and descriptive format for packaging representations of others' mental states, language may have allowed our ancestors to represent more complex beliefs and desires and higher degrees of recursion than is possible for nonverbal species. However, no existing data can speak to the scope of the capacity for recursion in great apes. Fourth, it remains unknown whether nonhumans (or even young children) are fully conscious of their representations of others' mental states (i.e., explicitly as opposed to implicitly representing them), in the way that human adults are. Finally, Tomasello et al. (2005) have argued that humans are unique in our ability to structure cooperative activities around shared goals and joint representations of the world, perhaps owing in part to uniquely cooperative and interdependent motivations (Bullinger, Melis, & Tomasello, 2011; Krupenye et al., 2018; MacLean & Hare, 2013; Melis & Tomasello, 2013; Rekers, Haun, & Tomasello, 2011). Tomasello (2018) has also argued that this uniquely human variety of cooperative action and thinking scaffolds our developing understanding of objective perspectives and our unique ability to coordinate and compare mental states with one another. Future work will prove essential in determining which of these capacities—and which others—are truly unique to humans and can account for the unrivaled complexity of our social minds and our social worlds.

Summary

Although future data will continue to refine our inferences, the present exercise in examining theory of mind from a phylogenetic perspective has allowed us to sketch out a rough timeline of key changes in social cognition throughout human and primate evolutionary history (Fig. 1). This exercise suggests that the roots of our social minds extend deep into primate phylogeny. Already 77 mya, the common ancestor of all living primates appeared capable of responding adaptively to social cues, such as body and facial orientation, in order to follow gaze to objects or events of interest, steal uncontested food, and presumably evade predators. However, their social understanding may have been limited to simple heuristic responses to coarse social cues; there is yet no evidence that strepsirrhine primates can integrate information about what another can perceive from multiple sensory modalities. It is unknown how richly this ancestor attributed agency, goals, and intentions to others; however, by 43 mya, the common ancestor of all living haplorhine primates appeared to interpret others' actions in terms of the motivations that underlie them. Perhaps at this point but certainly by 30 mya (when the catarrhine radiation began), our ancestors also possessed a more robust understanding of others' perspectives. They could track, in some sense, what others could see and hear and integrate these types of information to adeptly navigate social competition. They could also represent, in some sense, what others had previously seen and therefore knew. These capacities may reflect attribution of knowledge and ignorance or of awareness relations. By 16–20 mya (the common ancestor of the great apes), the evidence is even stronger that these representations were already mentalistic: apes use self-experience, perhaps through experience-projection, to predict perception-dependent and belief-based actions. Our great ape common ancestor could likely also represent simple forms of false beliefs, at least about the location of hidden objects, but the nature and diversity of these representations remain unknown. Human social cognition is additionally characterized by a number of capacities that, to date, have not been demonstrated in other species—such as full-blown propositional attitude psychology, level II perspective-taking, higher-level recursive mind-reading, shared intentionality, and explicit representations of others' mental states. This comparative view provides evidence that there has likely been a consistent elaboration of social cognition, and of the complexity of the computations that underlie it, throughout human and primate evolutionary history (Humphrey, 1976).

It is important to note once more that this attempt to reconstruct the ancestral states of primate cognition is based on highly incomplete data, in some instances with single species standing in for much larger clades. More inclusive comparisons across all primate groups—and targeting of particularly relevant unstudied species—will be essential for more completely charting the evolutionary history of our social minds (Nunn, 2011). Some biases may also pervade our existing data. First, the species that have shown the greatest abilities (e.g., chimpanzees) have also received the greatest sampling effort. Although it would be highly surprising if important differences did not exist between our closer relatives and our more distant ones, more equitable sampling will be important for determining exactly what they are. Second, some comparative tasks have been designed with humans in mind and may naturally be more intuitive to more similar species, like chimpanzees, while other tasks have been designed to maximize the ecological validity for particular species and may be less motivating for others (e.g., competitive tasks, which are highly motivating for chimpanzees but may instead be stressful for bonobos) (Wobber et al., 2010). Pursuing direct comparisons involving tasks that have been validated in some taxa as well as developing novel tasks that maximize motivation for target species will also be critical for accurately characterizing the similarities and nuanced differences between species (Krupenye, MacLean, et al., 2017). Finally, we must continue to pursue controlled efforts to isolate the specific cognitive mechanisms, and their underlying representations, that separate one species from another, including those that uniquely characterize human theory of mind. All of these investigations will prove fundamental to determining the evolutionary history of our social minds.

Identifying the Selective Pressures that Built Human Theory of Mind

Our phylogenetic approach has offered a window into the patterns of change that likely characterized the evolution of human theory of mind, but what drove those changes? Although there are several mechanisms of evolution, consistent directional change is most likely the result of natural selection acting either on the specific traits in question or traits with which they are associated. In the latter case, social cognitive evolution could be a byproduct of selection, e.g., for increased brain size or for something much less intuitively linked (in the case of complex pleiotropic effects). Understanding the cognitive, neural, and genetic bases of theory of mind will be necessary to confirm its history of selection, but given the clear utility of theory of mind it is worth surveying the pressures that may have played a role.

A trait can be favored by selection if its benefits to survival and reproduction outweigh its costs. Cognitive and neural traits are potentially metabolically costly and thus substantial consideration has been given to the ways in which changes in diet or gut morphology can potentially pay for these costs (Aiello & Wheeler, 1995; Pontzer et al., 2016; Wrangham, 2009). Diet might also directly drive different cognitive adaptations (the ecological intelligence hypothesis): for example, clumped and ephemeral resources like fruit place different demands on spatial cognition and memory than do uniformly distributed resources like leaves, and extractive foraging likely presents unique challenges for technical intelligence and social learning (Clutton-Brock & Harvey, 1980; Milton, 1988; Rosati, 2017). Phylogenetic comparisons suggest that species with more demanding ecologies indeed tend to have larger brains and also to perform better on cognitive tests of self-control (DeCasien, Williams, & Higham, 2017; MacLean et al., 2014; Powell, Isler, & Barton, 2017). Ecology likely plays a fundamental role in determining the constraints on and drivers of cognitive evolution, including of social cognitive traits.

However, while diet is clearly central to the evolutionary process, the most readily apparent functions of theory of mind are social. A variety of theoreticians have proposed that social cognition evolved in response to the demands of group living (the social intelligence hypothesis) (Byrne & Bates, 2007; Byrne & Whiten, 1988; Dunbar & Shultz, 2007; Humphrey, 1976; Jolly, 1966). Since social species must compete (or coordinate) with their own groupmates for reproductive opportunities, those that can outmaneuver their competitors should experience the greatest social and reproductive success—and, in a ratcheting process, selection should consistently favor cognitive skills that improve this social maneuvering. The finding from the first part of this paper that social cognition appears to have been consistently elaborated throughout primate and human evolution provides some evidence for such a ratcheting process.

Comparisons among distantly related taxa also suggest that sophisticated social cognitive traits have evolved convergently, multiple times, in clades that feature particularly complex and demanding social organizations, such as primates, whales and dolphins, elephants, and birds in the crow family (Emery & Clayton, 2004). Meanwhile, comparisons of closely related taxa (lemurs and also some birds) have shown that species living in more complex groups tend to perform better than their less social relatives on tasks that directly tap skills relevant to social life (Bond, Kamil, & Balda, 2003; MacLean, Merritt, & Brannon, 2008; MacLean, Sandel, Bray, et al., 2013; Sandel et al., 2011). These data suggest that while brain size generally may be more strongly tied to feeding ecology than sociality, specific social cognitive traits may have evolved in response to social pressures. However, to date, no large-scale comparisons have been performed, making it less clear whether these patterns would generalize across much larger taxonomic groups or replicate in studies that test more sophisticated theory of mind abilities.

Critically, the social intelligence hypothesis asserts that social cognition has been favored specifically because of the fitness benefits it provides to the most socially savvy. Observational studies have provided a large body of anecdotal evidence that apes and other primates use deceptive behavior, which could be underlain by theory of mind, to access food and mates and to avoid aggression (Byrne & Whiten, 1988; de Waal, 1982; Whiten & Byrne, 1988). Field and captive experiments also suggest, at the population level, that nonhuman primates can eavesdrop or use social information adaptively (Crockford, Wittig, Seyfarth, & Cheney, 2007; Hare et al., 2000; Wittig et al., 2014). However, little work has directly attempted to link individual differences in cognition to differential acquisition of proximate or ultimate benefits (Ashton, Ridley, Edwards, & Thornton, 2018), although, importantly, such variation in cognition appears to be heritable (Hopkins, Russell, & Schaeffer, 2014).

Noting the need for continued testing of the ecological and social intelligence hypotheses, existing work suggests that human theory of mind has likely evolved in response to a mosaic of pressures from both the social and physical worlds. Its evolution (and that of primates' large brains more broadly) has surely been shaped by the constraints imposed and relieved by dietary and metabolic adaptations. Meanwhile, specific social cognitive specializations likely stem from the demands of living in complex groups. In these contexts, theory of mind seems to serve many adaptive functions: for interpreting, predicting, and manipulating others' behavior, for cooperating and competing, and perhaps for communicating and (certainly in humans) teaching.

Synthesis and Future Directions

Theory of mind is at the heart of what makes us human, but phylogenetic comparisons evince deep evolutionary roots. From early primates that could respond adaptively to social cues, like body and facial orientation, to an ape common ancestor that may have understood simple beliefs, comparative research suggests that the social cognitive skills that define humans did not emerge overnight; they have steadily been elaborated throughout the last 77 million years of our evolutionary past. And yet a number of notable features of human theory of mind—like recursive mind-reading, explicit mental state understanding, and shared intentionality—may well have appeared in the last 6–9 million years, since our species diverged from the other apes. Our social minds are likely the product of selection for cooperating and competing with groupmates, additionally shaped by the constraints of feeding ecology.

In the 40 years that followed Premack and Woodruff's (1978) seminal investigations of chimpanzee theory of mind, we have learned a great deal about the evolution of human mentalizing. And still, unanswered questions abound. Future experimental work must endeavor to precisely specify the cognitive representations (and neural underpinnings) that support social cognition in each species, to understand what exactly separates humans from chimpanzees and bonobos and how the precursors of these traits are distributed across primate phylogeny. Broader comparisons will prove essential, both for accurately reconstructing the evolutionary history of human theory of mind and for testing hypotheses about its evolutionary function. Within species, we also need continued investigation of the genetic and environmental foundations of social cognition, of its ontogeny, and of the selective pressures that likely drove its evolution. Only with such a comprehensive approach will we ever be able to fully understand what exactly it is that makes us human and how we came to be.

Acknowledgements I am thankful to Josep Call, Brian Hare, Richard Byrne, and Andrew Whiten for discussions over the years on topics relevant to this paper. I am also grateful to the European Commission for financial support (Marie Skłodowska-Curie fellowship MENTALIZINGORIGINS). Finally, abundant thanks are owed to Michael Gilead and Kevin Ochsner, for their kind invitation to participate in this exciting compilation and for their enduring patience and support.

References

Adolphs, R. (2009). The social brain: Neural basis of social knowledge. *Annual Review of Psychology, 60*, 693–716. https://doi.org/10.1146/annurev.psych.60.110707.163514

Aiello, L. C., & Wheeler, P. (1995). The expensive-tissue hypothesis: The brain and the digestive system in human and primate evolution. *Current Anthropology, 36*(2), 199–221.

Anderson, J. R., Kuroshima, H., Takimoto, A., & Fujita, K. (2013). Third-party social evaluation of humans by monkeys. *Nature Communications, 4*, 1561. https://doi.org/10.1038/ncomms2495

Ashton, B. J., Ridley, A. R., Edwards, E. K., & Thornton, A. (2018). Cognitive performance is linked to group size and affects fitness in Australian magpies. *Nature, 554*(7692), 364–367. https://doi.org/10.1038/nature25503

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14*(3), 110–118.

Banaji, M. R., & Gelman, S. A. (Eds.). (2012). *Navigating the social world: What infants, children, and other species can teach us*. New York, NY: Oxford University Press.

Bettle, R., & Rosati, A. G. (2019). Flexible gaze-following in rhesus monkeys. *Animal Cognition, 22*, 673. https://doi.org/10.1007/s10071-019-01263-4

Bond, A. B., Kamil, A. C., & Balda, R. P. (2003). Social complexity and transitive inference in corvids. *Animal Behaviour, 65*, 479–487.

Boyd, R. (2006). The puzzle of human sociality. *Science, 314*, 1555–1556.

Brauer, J., Call, J., & Tomasello, M. (2005). All great ape species follow gaze to distant locations and around barriers. *Journal of Comparative Psychology, 119*(2), 145–154. https://doi.org/10.1037/0735-7036.119.2.145

Brauer, J., Call, J., & Tomasello, M. (2007). Chimpanzees really know what others can see in a competitive situation. *Animal Cognition, 10*(4), 439–448.

Brauer, J., Call, J., & Tomasello, M. (2008). Chimpanzees do not take into account what others can hear in a competitive situation. *Animal Cognition, 11*(1), 175–178. https://doi.org/10.1007/s10071-007-0097-0

Bray, J., Krupenye, C., & Hare, B. (2014). Ring-tailed lemurs (Lemur catta) exploit information about what others can see but not what they can hear. *Animal Cognition, 17*(3), 735–744. https://doi.org/10.1007/s10071-013-0705-0

Bullinger, A. F., Melis, A. P., & Tomasello, M. (2011). Chimpanzees, Pan troglodytes, prefer individual over collaborative strategies towards goals. *Animal Behaviour, 82*(5), 1135–1141. https://doi.org/10.1016/j.anbehav.2011.08.008

Burkart, J., Kupferberg, A., Glasauer, S., & van Schaik, C. (2012). Even simple forms of social learning rely on intention attribution in marmoset monkeys (Callithrix jacchus). *Journal of Comparative Psychology, 126*(2), 129–138. https://doi.org/10.1037/a0026025

Burkart, J. M., & Heschl, A. (2007). Understanding visual access in common marmosets, Callithrix jacchus: Perspective taking or behaviour reading? *Animal Behaviour, 73*, 457–469. https://doi.org/10.1016/j.anbehav.2006.05.019

Buttelmann, D., Buttelmann, F., Carpenter, M., Call, J., & Tomasello, M. (2017). Great apes distinguish true from false beliefs in an interactive helping task. *PLoS One, 12*(4), e0173793.

Buttelmann, D., Carpenter, M., Call, J., & Tomasello, M. (2007). Enculturated chimpanzees imitate rationally. *Developmental Science, 10*(4), F31–F38. https://doi.org/10.1111/j.1467-7687.2007.00630.x

Buttelmann, D., Schütte, S., Carpenter, M., Call, J., & Tomasello, M. (2012). Great apes infer others' goals based on context. *Animal Cognition, 15*, 1037–1053. https://doi.org/10.1007/s10071-012-0528-4

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language, 28*(5), 606–637.

Byrne, R. W., & Bates, L. A. (2007). Sociality, evolution and cognition. *Current Biology, 17*(16), R714–R723.

Byrne, R. W., & Whiten, A. W. (1988). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford: Clarendon Press.

Call, J., Hare, B., Carpenter, M., & Tomasello, M. (2004). 'Unwilling' versus 'unable': Chimpanzees' understanding of human intentional action. *Developmental Science, 7*(4), 488–498.

Call, J., & Tomasello, M. (1998). Distinguishing intentional from accidental actions in orangutans (Pongo pygmaeus), chimpanzees (Pan troglodytes), and human children (Homo sapiens). *Journal of Comparative Psychology, 112*(2), 192–206. https://doi.org/10.1037/0735-7036.112.2.192

Call, J., & Tomasello, M. (1999). A nonverbal false belief task: The performance of children and great apes. *Child Development, 70*(2), 381–395.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences, 12*(5), 187–192.

Canteloup, C., & Meunier, H. (2017). 'Unwilling' versus 'unable': Tonkean macaques' understanding of human goal-directed actions. *PeerJ, 5*, e3227. https://doi.org/10.7717/peerj.3227

Canteloup, C., Piraux, E., Poulin, N., & Meunier, H. (2016). Do Tonkean macaques (Macaca tonkeana) perceive what conspecifics do and do not see? *PeerJ, 4*, e1693. https://doi.org/10.7717/peerj.1693

Clutton-Brock, T. H., & Harvey, P. H. (1980). Primates, brains and ecology. *Journal of Zoology, 190*, 309–323.

Corballis, M. C. (2011). *The recursive mind: The origins of human language, thought, and civilization*. Princeton, NJ: Princeton University Press.

Crockford, C., Wittig, R. M., Mundry, R., & Zuberbuhler, K. (2012). Wild chimpanzees inform ignorant group members of danger. *Current Biology, 22*(2), 142–146. https://doi.org/10.1016/j.cub.2011.11.053

Crockford, C., Wittig, R. M., Seyfarth, R. M., & Cheney, D. L. (2007). Baboons eavesdrop to deduce mating opportunities. *Animal Behaviour, 73*(5), 885–890. https://doi.org/10.1016/j.anbehav.2006.10.016

Darwin, C. (1859). *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life*. London: John Murray.

de Waal, F. B. (1982). *Chimpanzee politics: Power and sex among apes*. New York, NY: Harper and Row.

DeCasien, A. R., Williams, S. A., & Higham, J. P. (2017). Primate brain size is predicted by diet but not sociality. *Nature Ecology & Evolution, 1*, 0112. https://doi.org/10.1038/s41559-017-0112

Defolie, C., Malassis, R., Serre, M., & Meunier, H. (2015). Tufted capuchins (Cebus apella) adapt their communicative behaviour to human's attentional states. *Animal Cognition, 18*(3), 747–755. https://doi.org/10.1007/s10071-015-0841-9

Drayton, L. A., & Santos, L. R. (2014). Capuchins' (Cebus apella) sensitivity to others' goal-directed actions in a helping context. *Animal Cognition, 17*(3), 689–700.

Drayton, L. A., & Santos, L. R. (2017). Do rhesus macaques, Macaca mulatta, understand what others know when gaze following? *Animal Behaviour, 134*, 193–199. https://doi.org/10.1016/j.anbehav.2017.10.016

Drayton, L. A., & Santos, L. R. (2018). What do monkeys know about others' knowledge? *Cognition, 170*, 201–208. https://doi.org/10.1016/j.cognition.2017.10.004

Dunbar, R. I., & Shultz, S. (2007). Evolution in the social brain. *Science, 317*(5843), 1344–1347.

Emery, N. J., & Clayton, N. S. (2004). The mentality of crows: Convergent evolution of intelligence in corvids and apes. *Science, 306*(5703), 1903–1907. https://doi.org/10.1126/science.1098410

Flavel, J. H., Everett, B. A., Croft, K., & Flavel, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. *Developmental Psychology, 17*(1), 99–103.

Flombaum, J. I., & Santos, L. R. (2005). Rhesus monkeys attribute perceptions to others. *Current Biology, 15*(5), 447–452.

Hare, B., Addessi, E., Call, J., Tomasello, M., & Visalberghi, E. (2003). Do capuchin monkeys, Cebus apella, know what conspecifics do and do not see? *Animal Behaviour, 65*, 131–142.

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour, 59*, 771–785.

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour, 61*(1), 139–151.

Hare, B., Call, J., & Tomasello, M. (2006). Chimpanzees deceive a human competitor by hiding. *Cognition, 101*(3), 495–514.

Herrmann, E., Call, J., Hernandez-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science, 317*(5843), 1360–1366.

Heyes, C. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences, 21*(1), 101–114. Discussion 115–148.

Hopkins, W. D., Russell, J. L., & Schaeffer, J. (2014). Chimpanzee intelligence is heritable. *Current Biology, 24*(14), 1649–1652. https://doi.org/10.1016/j.cub.2014.05.076

Horschler, D. J., Santos, L. R., & MacLean, E. L. (2019). Do non-human primates really represent others' ignorance? A test of the awareness relations hypothesis. *Cognition, 190*, 72–80. https://doi.org/10.1016/j.cognition.2019.04.012

Horton, K. E., & Caldwell, C. A. (2006). Visual co-orientation and expectations about attentional orientation in pileated gibbons (Hylobates pileatus). *Behavioural Processes, 72*(1), 65–73. https://doi.org/10.1016/j.beproc.2005.12.004

Humphrey, N. K. (1976). The social function of intellect. In P. Bateson & R. Hinde (Eds.), *Growing points in ethology* (pp. 303–317). Cambridge: Cambridge University Press.

Jolly, A. (1966). Lemur social behavior and primate intelligence. *Science, 153*, 501–506.

Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition, 109*(2), 224–234. https://doi.org/10.1016/j.cognition.2008.08.010

Kano, F., & Call, J. (2014). Great apes generate goal-based action predictions: An eye-tracking study. *Psychological Science, 25*(9), 1691–1698. https://doi.org/10.1177/0956797614536402

Kano, F., Krupenye, C., Hirata, S., & Call, J. (2017). Eye tracking uncovered great apes' ability to anticipate that other individuals will act according to false beliefs. *Communicative & Integrative Biology, 10*(2), e1299836. https://doi.org/10.1080/19420889.2017.1299836

Kano, F., Krupenye, C., Hirata, S., Call, J., & Tomasello, M. (2017). Submentalizing cannot explain belief-based action anticipation in apes. *Trends in Cognitive Sciences, 21*(9), 633–634. https://doi.org/10.1016/j.tics.2017.06.011

Kano, F., Krupenye, C., Hirata, S., Tomonaga, M., & Call, J. (2019). Great apes use self-experience to anticipate an agent's action in a false belief test. *Proceedings of the National Academy of Sciences of the United States of America, 116*, 20904.

Karg, K., Schmelz, M., Call, J., & Tomasello, M. (2015a). Chimpanzees strategically manipulate what others can see. *Animal Cognition, 18*(5), 1069–1076. https://doi.org/10.1007/s10071-015-0875-z

Karg, K., Schmelz, M., Call, J., & Tomasello, M. (2015b). The goggles experiment: Can chimpanzees use self-experience to infer what a competitor can see? *Animal Behaviour, 105*, 211–221. https://doi.org/10.1016/j.anbehav.2015.04.028

Karg, K., Schmelz, M., Call, J., & Tomasello, M. (2016). Differing views: Can chimpanzees do Level 2 perspective-taking? *Animal Cognition, 19*(3), 555–564. https://doi.org/10.1007/s10071-016-0956-7

Kovacs, A. M., Teglas, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science, 330*(6012), 1830–1834. https://doi.org/10.1126/science.1190792

Krachun, C., Carpenter, M., Call, J., & Tomasello, M. (2009). A competitive nonverbal false belief task for children and apes. *Developmental Science, 12*(4), 521–535. https://doi.org/10.1111/j.1467-7687.2008.00793.x

Krachun, C., Carpenter, M., Call, J., & Tomasello, M. (2010). A new change-of-contents false belief test: Children and chimpanzees compared. *International Journal of Comparative Psychology, 23*, 145–165.

Krupenye, C., & Call, J. (2019). Theory of mind in animals: Current and future directions. *Wiley Interdisciplinary Reviews: Cognitive Science, 10*, e1503. https://doi.org/10.1002/wcs.1503

Krupenye, C., & Hare, B. (2018). Bonobos prefer individuals that hinder others over those that help. *Current Biology, 28*(2), 280–286.e285. https://doi.org/10.1016/j.cub.2017.11.061

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science, 354*(6308), 110–114.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2017). A test of the submentalizing hypothesis: Apes' performance in a false belief task inanimate control. *Communicative & Integrative Biology, 10*(4), e1343771. https://doi.org/10.1080/19420889.2017.1343771

Krupenye, C., MacLean, E., & Hare, B. (2017). Does the bonobo have a (chimpanzee-like) theory of mind? In B. Hare & S. Yamamoto (Eds.), *Bonobos: Unique in mind, brain and behavior*. Oxford: Oxford University Press.

Krupenye, C., Tan, J., & Hare, B. (2018). Bonobos voluntarily hand food to others but not toys or tools. *Proceedings of the Biological Sciences, 285*(1886), 20181536. https://doi.org/10.1098/rspb.2018.1536

Kupferberg, A., Glasauer, S., & Burkart, J. M. (2013). Do robots have goals? How agent cues influence action understanding in non-human primates. *Behavioural Brain Research, 246*, 47–54. https://doi.org/10.1016/j.bbr.2013.01.047

Liebal, K., & Kaminski, J. (2012). Gibbons (Hylobates pileatus, H. moloch, H. lar, Symphalangus syndactylus) follow human gaze, but do not take the visual perspective of others. *Animal Cognition, 15*, 1211–1216. https://doi.org/10.1007/s10071-012-0543-5

Lorincz, E. N., Jellema, T., Gómez, J. C., Barraclough, N., Xiao, D., & Perrett, D. I. (2005). Do monkeys understand actions and minds of others? Studies of single cells and eye movements. In S. Dehaene, J. R. Duhamel, M. D. Hauser, & G. Rizzolatti (Eds.), *From monkey brain to human brain: A Fyssen Foundation Symposium* (pp. 189–210). Cambridge, MA: MIT Press.

Lurz, R. (2009). If chimpanzees are mindreaders, could behavioral science tell? Toward a solution of the logical problem. *Philosophical Psychology, 22*(3), 305–328. https://doi.org/10.1080/09515080902970673

Lurz, R., Krachun, C., Mahovetz, L., Wilson, M. J. G., & Hopkins, W. (2018). Chimpanzees gesture to humans in mirrors: Using reflection to dissociate seeing from line of gaze. *Animal Behaviour, 135*, 239–249. https://doi.org/10.1016/j.anbehav.2017.11.014

MacLean, E. L., & Hare, B. (2012). Bonobos and chimpanzees infer the target of another's attention. *Animal Behaviour, 83*(2), 345–353. https://doi.org/10.1016/j.anbehav.2011.10.026

MacLean, E. L., & Hare, B. (2013). Spontaneous triadic engagement in bonobos (Pan paniscus) and chimpanzees (Pan troglodytes). *Journal of Comparative Psychology, 127*(3), 245–255. https://doi.org/10.1037/a0030935

MacLean, E. L., Hare, B., Nunn, C. L., Addessi, E., Amici, F., Anderson, R. C., … Zhao, Y. (2014). The evolution of self-control. *Proceedings of the National Academy of Sciences of the United States of America, 111*(20), E2140–E2148.

MacLean, E. L., Matthews, L., Hare, B., Nunn, C. L., Anderson, R. C., Aureli, F., … Wobber, V. (2012). How does cognition evolve? Phylogenetic comparative psychology. *Animal Cognition, 15*, 223–238.

MacLean, E. L., Merritt, D. J., & Brannon, E. M. (2008). Social complexity predicts transitive reasoning in prosimian primates. *Animal Behaviour, 76*(2), 479–486. https://doi.org/10.1016/j.anbehav.2008.01.025

MacLean, E. L., Sandel, A. A., Bray, J., Oldenkamp, R. E., Reddy, R. B., & Hare, B. A. (2013). Group size predicts social but not nonsocial cognition in lemurs. *PLoS One, 8*(6), e66359. https://doi.org/10.1371/journal.pone.0066359

MacLean, E. L., Sandel, A. A., Reddy, R., Bray, J., Oldenkamp, R., & Hare, B. (2013). Group size predicts social, but not nonsocial cognition in lemurs. *PLoS One, 8*(6), e66359. https://doi.org/10.1371/journal.pone.0066359

Marticorena, D. C. W., Ruiz, A. M., Mukerji, C., Goddu, A., & Santos, L. R. (2011). Monkeys represent others' knowledge but not their beliefs. *Developmental Science, 14*(6), 1406–1416. https://doi.org/10.1111/j.1467-7687.2011.01085.x

Martin, A., & Santos, L. R. (2014). The origins of belief representation: Monkeys fail to automatically represent others' beliefs. *Cognition, 130*(3), 300–308. https://doi.org/10.1016/j.cognition.2013.11.016

Martin, A., & Santos, L. R. (2016). What cognitive representations support primate theory of mind? *Trends in Cognitive Sciences, 20*(5), 375–382. https://doi.org/10.1016/j.tics.2016.03.005

Melis, A. P., Call, J., & Tomasello, M. (2006). Chimpanzees (Pan troglodytes) conceal visual and auditory information from others. *Journal of Comparative Psychology, 120*(2), 154–162. https://doi.org/10.1037/0735-7036.120.2.154

Melis, A. P., & Tomasello, M. (2013). Chimpanzees' (Pan troglodytes) strategic helping in a collaborative task. *Biology Letters, 9*(2), 20130009. https://doi.org/10.1098/rsbl.2013.0009

Meltzoff, A. N., & Brooks, R. (2008). Self-experience as a mechanism for learning about others: A training study in social cognition. *Developmental Psychology, 44*(5), 1257–1265. https://doi.org/10.1037/a0012888

Meunier, H. (2017). Do monkeys have a theory of mind? How to answer the question? *Neuroscience and Biobehavioral Reviews, 82*, 110–123. https://doi.org/10.1016/j.neubiorev.2016.11.007

Milton, K. (1988). Foraging behavior and the evolution of primate cognition. In A. Whiten & R. Byrne (Eds.), *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes and humans* (pp. 285–305). Oxford: Oxford University Press.

Moll, H., & Meltzoff, A. N. (2011). How does it look? Level 2 perspective-taking at 36 months of age. *Child Development, 82*(2), 661–673. https://doi.org/10.1111/j.1467-8624.2010.01571.x

Muller, M. N., Wrangham, R. W., & Pilbeam, D. R. (2017). *Chimpanzees and human evolution*. Cambridge, MA: Harvard University Press.

Myowa-Yamakoshi, M., & Matsuzawa, T. (2000). Imitation of intentional manipulatory actions in chimpanzees. *Journal of Comparative Psychology, 114*(4), 381–391.

Nunn, C. L. (2011). *The comparative approach in evolutionary anthropology and biology*. Chicago, IL: The University of Chicago Press.

O'Grady, C., Kliesch, C., Smith, K., & Scott-Phillips, T. C. (2015). The ease and extent of recursive mindreading, across implicit and explicit tasks. *Evolution and Human Behavior, 36*(4), 313–322. https://doi.org/10.1016/j.evolhumbehav.2015.01.004

Okamoto-Barth, S., Call, J., & Tomasello, M. (2007). Great apes' understanding of other individuals' line of sight. *Psychological Science, 18*(5), 462–468.

Onishi, K. H., & Bailargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255–258.

Penn, D., & Povinelli, D. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society* *of London. Series B, Biological Sciences, 362*(1480), 731–744. https://doi.org/10.1098/rstb.2006.2023. J577253MH830J082 [pii].

Phillips, W., Barnes, J. L., Mahajan, N., Yamaguchi, M., & Santos, L. R. (2009). 'Unwilling' versus 'unable': Capuchin monkeys' (Cebus apella) understanding of human intentional action. *Developmental Science, 12*(6), 938–945. https://doi.org/10.1111/J.1467-7687.2009.00840.X

Pontzer, H., Brown, M. H., Raichlen, D. A., Dunsworth, H., Hare, B., Walker, K., … Ross, S. R. (2016). Metabolic acceleration and the evolution of human brain size and life history. *Nature, 533*, 390–392. https://doi.org/10.1038/nature17654

Povinelli, D. J., & Eddy, T. J. (1996). Chimpanzees: Join visual attention. *Psychological Science, 7*(3), 129–135.

Povinelli, D. J., & Vonk, J. (2003). Chimpanzee minds: Suspiciously human? *Trends in Cognitive Sciences, 7*(4), 157–160. S1364661303000536 [pii].

Powell, L. E., Isler, K., & Barton, R. A. (2017). Re-evaluating the link between brain size and behavioural ecology in primates. *Proceedings of the Biological Sciences, 284*(1865), 20171765. https://doi.org/10.1098/rspb.2017.1765

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind. *Behavioral and Brain Sciences, 1*(4), 515–526.

Rekers, Y., Haun, D., & Tomasello, M. (2011). Children, but not chimpanzees, prefer to collaborate. *Current Biology, 21*, 1756.

Rochat, M. J., Serra, E., Fadiga, L., & Gallese, V. (2008). The evolution of social cognition: Goal familiarity shapes monkeys' action understanding. *Current Biology, 18*(3), 227–232. https://doi.org/10.1016/j.cub.2007.12.021

Rosati, A. G. (2017). Foraging cognition: Reviving the ecological intelligence hypothesis. *Trends in Cognitive Sciences, 21*(9), 691–702. https://doi.org/10.1016/j.tics.2017.05.011

Rosati, A. G., & Hare, B. (2009). Looking past the model species: Diversity in gaze-following skills across primates. *Current Opinion in Neurobiology, 19*(1), 45–51.

Ruiz, A., Gomez, J. C., Roeder, J. J., & Byrne, R. W. (2009). Gaze following and gaze priming in lemurs. *Animal Cognition, 12*, 427.

Sandel, A. A., MacLean, E., & Hare, B. (2011). Evidence from four lemur species that ringtailed lemur social cognition converges with that of haplorhine primates. *Animal Behaviour, 81*, 925–931.

Santos, L. R., Nissen, A. G., & Ferrugia, J. A. (2006). Rhesus monkeys, Macaca mulatta, know what others can and cannot hear. *Animal Behaviour, 71*(5), 1175–1181. https://doi.org/10.1016/j.anbehav.2005.10.007

Senju, A., Southgate, V., Snape, C., Leonard, M., & Csibra, G. (2011). Do 18-month-olds really attribute mental states to others? A critical test. *Psychological Science, 22*(7), 878–880. https://doi.org/10.1177/0956797611411584

Shepherd, S. V., & Platt, M. L. (2008). Spontaneous social orienting and gaze following in ringtailed lemurs (Lemur catta). *Animal Cognition, 11*(1), 13–20. https://doi.org/10.1007/s10071-007-0083-6

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18*(7), 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Steiper, M. E., & Young, N. M. (2006). Primate molecular divergence dates. *Molecular Phylogenetics and Evolution, 41*(2), 384–394. https://doi.org/10.1016/j.ympev.2006.05.021

Sterelny, K. (2019). The origins of multi-level society. *Topoi*. https://doi.org/10.1007/s11245-019-09666-1

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für Tierpsychologie, 20*, 410–433.

Tomasello, M. (2018). How children come to understand false beliefs: A shared intentionality account. *Proceedings of the National Academy of Sciences of the United States of America, 115*(34), 8491–8498. https://doi.org/10.1073/pnas.1804761115

Tomasello, M., Call, J., & Hare, B. (1998). Five primate species follow the visual gaze of conspecifics. *Animal Behaviour, 55*, 1063–1069.

Tomasello, M., & Carpenter, M. (2005). The emergence of social cognition in three young chimpanzees. *Monographs of the Society for Research in Child Development, 70*(279), vii.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 675.

Tomasello, M., & Haberl, K. (2003). Understanding attention: 12-and 18-month-olds know what is new for other persons. *Developmental Psychology, 39*(5), 906–912.

Tomasello, M., Hare, B., Lehmann, H., & Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: The cooperative eye hypothesis. *Journal of Human Evolution, 52*(3), 314–320.

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two key steps in the evolution of human cooperation. *Current Anthropology, 53*(6), 673–692. https://doi.org/10.1086/668207

Tremblay, S., Sharika, K. M., & Platt, M. L. (2017). Social decision-making and the brain: A comparative perspective. *Trends in Cognitive Sciences, 21*(4), 265–276. https://doi.org/10.1016/j.tics.2017.01.007

Warneken, F., Hare, B., Melis, A. P., Hanus, D., & Tomasello, M. (2007). Spontaneous altruism by chimpanzees and young children. *PLoS Biology, 5*(7), e184.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science, 311*(5765), 1301–1303. https://doi.org/10.1126/science.1121448

Whiten, A. (2013). Humans are not alone in computing how others see the world. *Animal Behaviour, 86*(2), 213–221. https://doi.org/10.1016/j.anbehav.2013.04.021

Whiten, A., & Byrne, R. W. (1988). Tactical deception in primates. *The Behavioral and Brain Sciences, 11*(02), 233–273. https://doi.org/10.1017/s0140525x00049682

Wittig, R. M., Crockford, C., Langergraber, K. E., & Zuberbuhler, K. (2014). Triadic social interactions operate across time: A field experiment with wild chimpanzees. *Proceedings of the Biological Sciences, 281*(1779), 20133155. https://doi.org/10.1098/rspb.2013.3155

Wobber, V., Hare, B., Maboto, J., Lipson, S., Wrangham, R., & Ellison, P. T. (2010). Differential changes in steroid hormones before competition in bonobos and chimpanzees. *Proceedings of the National Academy of Sciences of the United States of America, 107*(28), 12457–12462.

Wrangham, R. W. (2009). *Catching fire: How cooking made us human*. New York, NY: Basic Books.

Yamamoto, S., Humle, T., & Tanaka, M. (2012). Chimpanzees' flexible targeted helping based on an understanding of conspecifics' goals. *Proceedings of the National Academy of Sciences of the United States of America, 109*(9), 3588–3592.

Mentalizing in Nonhuman Primates

Alyssa M. Arre and Laurie R. Santos

Humans effortlessly infer the mental states of other agents, spontaneously making swift and accurate predictions about how others will act based on these inferences. These theory of mind capacities are early emerging, with human infants beginning to make accurate predictions about how other agents should act based on the other agent's mental states within the first 2 years of life (Buttelmann, Carpenter, & Tomasello, 2009; Helming, Strickland, & Jacob, 2014; Luo, 2011; Onishi & Baillargeon, 2005; Scott & Baillargeon, 2017; Sodian, 2011). The end result is a sophisticated set of social cognitive abilities that make the human species an outlier in most domains, especially in terms of our hyper-collaboration and unique cultural evolution (Seed & Tomasello, 2010; Tomasello, 2000; Tomasello, Carpenter, Call, Behne, & Moll, 2005).

But are humans alone in our capacity to represent the minds of others? Or do we share at least some of our mentalizing capacities with our closest living relatives, the nonhuman primates (hereafter, primates). Like humans, most primate species live in large social groups, and thus it would be adaptive as for socially living primates to share many of the same theory of mind abilities of humans (Byrne & Bates, 2010). For the past four decades, researchers have devoted much empirical effort to testing whether primates share human-like mentalizing abilities (see reviews in Call & Tomasello, 2008; Krupenye & Call, 2019; Rosati, Santos, & Hare, 2010). Here, we explore what this research has taught us a date, with the goal of providing a unified account of what primates do and do not understand about other agents' mental states.

A. M. Arre (✉) · L. R. Santos
Department of Psychology, Yale University, New Haven, CT, USA
e-mail: alyssa.arre@yale.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_7

Four Decades of Primate Theory of Mind

Over 40 years ago, Premack and Woodruff (1978) were the first to ask whether chimpanzees possess a theory of mind. Their seminal work attempted to investigate the mentalizing abilities of a single chimpanzee, Sarah, using a series of tasks involving videos of a human facing a problem (e.g., being stuck in a locked cage) and subsequent photographs of possible solutions to the problem. Sarah chose the photograph that depicted the correct solution to the problem, which Premack and Woodruff interpreted as evidence that Sarah recognized both the experimenter's mental state (i.e., his intentions), as well as what was needed in order to fulfill the experimenter's goal.

Although many have debated Premack & Woodruff's initial interpretation of these findings (e.g., Dennett, 1978; Pylyshyn, 1978), their seminal paper launched several decades of work on the development of these abilities across human infancy and childhood (e.g., Astington & Gopnik, 1991; Flavell, 1999; Gopnik & Wellman, 1992; Wellman & Woolley, 1990; Wellman, Cross & Watson, 2001), which has revealed much about how mentalizing arises over the lifecourse and the different component processes that human children develop in order to represent the mental states of others (e.g., Wellman & Liu, 2004; Scott & Baillargeon, 2017; Sodian, 2011). Simultaneously, Premack and Woodruff launched a long line of work investigating the mentalizing capacities of nonhuman primates as well (see reviews in Call & Tomasello, 2011; Krupenye & Call, 2019; Rosati et al., 2010). Here, we explore what that work has shown about primate mentalizing. Throughout, we'll argue that understanding the combined (and often confusing) pattern of primates' successes and failures on these tasks will help us better understand not just how primates think about other minds but what primates can tell us about human mentalizing representations as well.

Representing that Agents Are Aware

Much of our human theory of mind reasoning involves tracking what other individuals are *aware* of: we track whether others share the same information we have, whether someone has noticed our indiscretions, and whether we need to inform our friends of new heretofore unknown gossip. Much research in human development has shown that this capacity to track what others are aware of emerges surprisingly early in human development. Human infants are able to track what others are looking at (Hains & Muir, 1996; Hood, Willen, & Driver, 1998; Symons, Hains, & Muir, 1998) or have seen in the past (Luo & Johnson, 2009; Song & Baillargeon, 2007; Song, Baillargeon, & Fisher, 2005) and use this information to make informed predictions about how an agent will act in the future (e.g., Onishi & Baillargeon, 2005). But do nonhuman primates share this capacity to represent awareness in others? As we review below, a number of studies using a variety of different kinds of tasks appears to converge on clear evidence that primates share this awareness representation capacity with humans (Crockford, Wittig, Mundry, & Zuberbühler, 2012; Flombaum & Santos, 2005; Hare, Call, Agnetta, & Tomasello, 2000; Hare, Call, & Tomasello, 2001, 2006; Hattori, Kano, & Tomonaga, 2010; Hirata & Matsuzawa, 2001; Hostetter, Russell, Freeman, & Hopkins, 2007; Kaminski et al., 2008; Karg, Schmelz, Call, & Tomasello, 2015; Marticorena et al., 2011; Melis, Call, & Tomasello, 2006; Santos, Nissen, & Ferrugia, 2006; Schmelz, Call, & Tomasello, 2011).

Gaze Following Tasks

Eyes serve as the window to what other social agents' are aware of. As such, researchers have long considered attention to others' faces and direction of gaze to be a foundational skill needed for a rich understanding of others' awareness (Wellman, 2011). For this reason, much of the early empirical work testing primates understanding of others' awareness began by exploring whether primates are able to follow the gaze of another individual. This work has shown that gaze following is widespread across the primate order, with many species of primates naturally following the gaze of conspecifics and human experimenters (apes: Povinelli & Eddy, 1996; Tomasello, Call, & Hare, 1998; Old World monkeys: Emery, Lorincz, Perrett, Oram, & Baker, 1997; Tomasello et al., 1998; New World monkeys: Burkart & Heschl, 2006; Neiworth, Burman, Basile, & Lickteig, 2002; prosimians: Sandel, MacLean, & Hare, 2011; Shepherd & Platt, 2008; Ruiz, Gómez, Roeder, & Byrne, 2009; Botting et al. (2011), for a review, see Rosati & Hare, 2009). Nevertheless, as many scholars have pointed out (Friesen & Kingstone, 1998; Penn & Povinelli, 2007), success on a gaze following task may not be indicative of a sophisticated understanding of others' awareness and may instead be nothing more than a reflexive process (i.e., co-orienting without any sort of meaningful representation of the content of the gaze). For this reason, primate researchers have developed more complex gaze following tasks, ones that require subjects to follow gaze towards a specific target object, often around a barrier or through a window. Such new geometric gaze following tasks require subjects to recognize that there is a *referent* to an agent's gaze (i.e., that the agent is *aware of something*, not just looking in some direction). While early work with great apes demonstrated our closest living relatives shared our human-like ability to geometrically gaze follow (Bräuer, Call, & Tomasello, 2005; MacLean & Hare, 2012; Okamoto-Barth, Call, & Tomasello, 2007; Tomasello, Hare, & Agnetta, 1999), more recent work has found evidence for this capacity in more distantly related monkeys as well (Amici, Aureli, Visalberghi, & Call, 2009; Bettle & Rosati, 2019).

In an even more direct use of gaze following to test whether primates represent others' awareness, MacLean and Hare (2012) tested chimpanzees on a modified gaze task where they directly varied whether the agent was aware of a target object. Specifically, the researchers varied whether the object a surprised agent looked at was novel (a toy she was unaware of and had never seen before) or familiar (one she had just seen, and thus should be uninterested in). If subjects understand that individuals rarely get surprised by objects they are aware of, then they should assume that the target object is the referent of the agent's gaze in the novel condition but not the familiar condition. Chimpanzees (and in later studies, rhesus monkeys, see Drayton & Santos, 2017) showed just this pattern of performance—they assumed that the agent was looking at the object in the ignorant condition (and just followed her gaze to the object) but assumed the agent must have a different referent when she was familiar with the object (and thus tracked her gaze beyond the object and out into open space). In this way, gaze following work shows that several primate species seem to gaze follow based on what a human experimenter is aware of and is not solely a reflexive reaction.

Competitive Tasks

Researchers have also observed evidence of primates' understanding of others' awareness using competitive tasks, where subjects must take into account what other agents are aware of when competing for resources in naturalistic situations (Bräuer et al., 2006; Flombaum & Santos, 2005; Hare et al., 2000, 2001; Kaminski et al., 2008; Santos et al., 2006, see Hare & Tomasello, 2004 and Lyons & Santos, 2006 for a review of these competitive tasks).

In the first of such tasks, Hare et al. (2000) placed dominant and subordinate chimpanzees into competition over two food rewards. The subordinate subject was able to see the position of both food rewards, while the dominant individual could only see one. If subordinates are able to track what dominant chimpanzees are aware of, then they should be more likely to steal food that the dominant individuals can't see. Hare et al. (2000) found that subordinate chimpanzees performed well on this task, successfully using information about what others were aware of during competition (see also Hare et al., 2001; Kaminski, Call, & Tomasello, 2008). Flombaum and Santos (2005) used a similar competitive design with free-ranging rhesus monkeys; they found that monkeys preferentially stole food from a human agent whose visual access was obscured (e.g., by turning away from the food, occluding the experimenter's face) but not from an agent who was aware of the location of the food.

Santos et al. (2006) tested the same macaque population on an auditory version of this stealing task. Monkeys could steal from one of two boxes in front of a human competitor: a *silent* box that opened and closed quietly or a *noisy* box covered in jingle bells that made noise when touched. If monkeys understand that noise can cause an unaware agent become aware, then they should preferentially steal from the silent box. Monkey showed just this pattern of performance, suggesting they are able to incorporate an additional sensory modality (audition) when considering another agent's awareness state (see also Melis et al., 2006 for similar results in apes).

Looking Time Tasks

Evidence that primates can accurately predict how another agent will act based on their awareness also comes from violation of expectation tasks (Drayton & Santos, 2018; Arre, Stumph, & Santos, 2021; Horschler, Santos, & MacLean, 2019; Arre, Clark, & Santos, 2020; Marticorena et al., 2011; Martin & Santos, 2014; for a review, see Drayton & Santos, 2016). Modeled off similar tasks used in human infants (e.g., Onishi & Baillargeon, 2005), the logic of these studies is that subject will look longer at events that violate their expectations compared to control scenes where no expectations are violated. In one study (Marticorena et al., 2011), rhesus monkeys watched as a human agent saw an object slide into one of two boxes. The agent then performed one of two actions: she reached into the box with the object (which should be consistent with her awareness that the object is in the box) or she reached into the empty box (and thus acted inconsistently with respect to her awareness of the object's location). Martincorena and colleagues found that macaques looked longer at the unexpected condition in which the experimenter reached towards the empty box, suggesting that they expect agents to act in accord with their visual awareness.

In another example, Drayton and Santos (2018) used a rotational displacement display to test whether subjects expected an agent to update their awareness of where an object was located. In this task, subjects saw an experimenter hide a piece of food in one of two boxes and then both the subject and the experimenter watched the two boxes rotate 180°. If monkeys understand that people can flexibly update what they're aware of, then they should expect the agent to reach in the object's newly displaced location rather than the original location. Monkeys showed just this expectation, looking longer when the agent reached in the original box. Importantly, rhesus monkeys only expected the agent to know the location of the hidden object when she herself had witnessed the difficult rotational displacement. These results suggest that primates recognize that awareness comes from perceptual access to a relevant event, and further, that primates' perceptual awareness representations are malleable when the context calls for it.

Taken together, a number of comparative studies using a variety of different methodologies presents clear evidence that primates are able to track what others are aware of. In addition, this work also suggests that primates' representations of other agents' perceptual awareness are multimodal and flexible, suggesting a robust mechanism that can update online as the subject gains more information about the other agent. Moreover, there is clear evidence that primates are further able to use these awareness representations both to make predictions about how agents will act in the future, and to determine their own best course of action in the social world.

Representing Others' Beliefs

Adult humans spend considerable time thinking about how another agent's subjective (and sometimes incorrect) worldview might influence their behavior. This striking ability to represent other individuals' *belief* states in this way requires that we generate both our own representation about the world, as well as another agent's potentially untrue or unfounded belief. Perhaps surprisingly given the complexity of these representations, researchers have observed such belief state representational abilities in infants younger than 15 months (for a review, see Scott & Baillargeon, 2017, although see Baillargeon et al., 2018, Poulin-Dubois et al., 2018, and Powell et al., 2018 which review some controversy about these findings). Using a nonverbal looking time version of the famous Sally Anne Task (Wimmer & Perner, 1983), Onishi and Baillargeon (2005) tested whether infant make predictions about how an experimenter should act based on what she believes. Fifteen-month-olds watched as a human agent witnessed a toy disappear into one of two locations. The agent then had her visual perspective occluded so that she could not see the presentation stage. While the agent was occluded, the infant saw the toy then move from its original location to the opposite location. Onishi and Baillargeon found that 15-month-olds looked longer when the agent reached into the *actual* location of the toy. In this way, infants seemed to expect that the agent should falsely believe that the object was in the first original location. Infants as young as 10 months old (Luo, 2011) have shown success on this and related tests of false beliefs (see review in Scott & Baillargeon, 2017), but what about primates? Do they also successfully represent others' beliefs?

Competitive Tasks

The earliest work exploring false belief capacities in primates tested these abilities in the context of a competitive food task. Kaminski et al. (2008) tested two chimpanzees on a turn-taking game in which they manipulated how much information each subject had about the content of three buckets on the table. Specifically, the researchers tested whether one chimpanzee (the subject) would change their choice behavior in response to what a competitor chimpanzee was aware of (Study 1) and believed (Study 2). In the first study, subjects watched as the competitor chimpanzee either saw or didn't see a high-quality food reward being hidden. Subjects then had a choice between that high-quality food reward or a low-quality alternative. Critically, the experimenters varied *when* subjects made their choice. When the competitor chose before the subject (and thus was likely to have already taken the high-quality food if they were aware of where it was), subjects selectively chose the low-quality option, but only if the competitor witnessed the baiting. Study 2 then built on this original design but included a false belief condition. Subjects in Study 2 failed to represent the false belief of the competitor, and instead, treated the agent as if they had no information about the content of the buckets at all. In this and other competitive studies (e.g., Krachun, Carpenter, Call, & Tomasello, 2009), primates fail to use others' false beliefs to successfully outcompete their foes.

Interactive Helping Tasks

Researchers have also attempted to test primate false belief understanding using interactive helping tasks, a method originally designed for use with human infants (Buttelmann et al., 2009). In the original infant version of this study, Buttelmann et al. (2009) allowed 16-, 18-, and 30-month-olds to watch as an agent placed an object in a box (A) which was then moved by another experimenter to a second box (B). Buttelmann and colleagues varied whether the agent saw the object's movement from box A to box B. The agent then intentionally tried to open box A. In the case where the agent had a true belief and therefore knew what was in both boxes, helping the agent achieve his goal would involve opening the box the agent was trying to open (box A). But when the agent had a false belief, helping to open box is no longer the best way to fulfill the agent's goal; instead, infants should selectively help the experimenter by opening box B, the one with the object actually inside. Infants show just this pattern of performance, helping the experimenter open box B on the false belief condition (77% of participants opening box B) but critically showing the opposite pattern of performance on the true belief condition (only 29% of participants opening box B).

Buttelmann and colleagues (2017) adapted this same task for use with great apes. Although apes mirrored infants' performance on the false belief condition (76% of trials opened box B), they showed a very different pattern of performance on the true belief condition, failing to distinguish between either of the two boxes (53% of trials opening box B). Although this result has been interpreted by some as evidence that apes distinguish between true and false belief states (Buttelmann et al., 2017), the difference in apes' performance from that of human infants raises doubts that primates interpret this task in the same way as humans do, and thus has made many scholars worry that this experiment cannot provide robust evidence that primates track others' beliefs.

Looking Time Studies

Researchers have also tested primate false belief understanding using expectancy violation looking time studies, including some measures borrowed from classic tests of theory of mind in human infants (e.g., Onishi & Baillargeon, 2005). Marticorena and colleagues (2011) presented rhesus monkeys with a scene in which an agent watched an object move between two possible locations. After the agent saw the object enter one of the two boxes, her visual perspective was blocked and the subject monkey alone watched as the object switched locations. If subjects represent that the agent now has a false belief, they should be surprised and subsequently look longer when the agent reaches into the box containing the object than when she reaches towards the box where she last saw the object (i.e., acting consistently with her false belief). In contrast to the performance of human infants (Onishi & Baillargeon, 2005), monkeys seem to show no prediction about where the experimenter will reach, looking for the same duration of time when she reaches to either box. Other looking time studies of false beliefs (Martin & Santos, 2014) have found similar failures when primates are allowed to automatically encode others' false beliefs (see Kovács et al. 2010 for a human infant version of this study). Taken together, the looking time studies match what has been observed in other kinds of primate false belief tasks—to date, there is no evidence that primates track what others believe. Indeed, when an agent has a false belief, primates for the most part seem to have no prediction about how that agent will behave.

Anticipatory Looking Methods

The single study to date providing positive evidence that apes may track others' false beliefs used anticipatory looking as a dependent measure (Krupenye, Kano, Hirata, Call, & Tomasello, 2016). In this task, apes watched videos of a human in an ape suit moving between different locations (or moving an object to different locations). Krupenye and colleagues then manipulated the amount of information a second agent in the video knew about the location of the first agent (or object). Critically, when the second agent in the video possessed a false belief about the location of the first agent (or the object), subjects made more anticipatory looks to the location where the second agent falsely believed the first agent (or object) to be. The researchers took this pattern of performance as evidence that apes indeed have false belief representational abilities, but that these abilities are fragile, elicited only in complex social situations and requiring a novel methodology to tap into the appropriate behavioral response (Krupenye et al., 2016; Krupenye & Call, 2019).

Conclusions

Although apes' performance on one single anticipatory looking task (Krupenye et al., 2016) has been interpreted by some as strong evidence that chimpanzees may understand the false beliefs of others (Krupenye & Call, 2019, see Krupenye, 2020, this volume), several scholars have criticized the paper on the grounds of small sample sizes and other methodological issues (for problems on the replicability of anticipatory looking methods for testing false beliefs in human development, see Baillargeon et al., 2018). Indeed, one co-author of the Krupenye et al. (2016) paper remains skeptical of that paper's interpretation, noting that this single piece of evidence has only "changed our conclusion somewhat … but in many studies, they still do not make behavioral decisions based on others' beliefs." (Tomasello, 2018, p. 180).

Apart from a single published report, primates have tended to perform poorly (Call & Tomasello, 1999; Kaminski et al., 2008; Krachun et al., 2009; O'Connell &

Dunbar, 2003) and differently than human infants (Buttelmann et al., 2017) on tests of false beliefs, even when tested with automatic processing tasks (Martin & Santos, 2014) and looking paradigms that require minimal task demands (Marticorena et al., 2011). Based on the preponderance of the evidence to date, we argue there is still relatively little reason to suspect that most primates successfully represent others' beliefs, at least in the way humans do.

Representing Unawareness in Others

While much of the existing comparative research tests whether primates understand what others are aware of and believe, there is less work specifically examining whether primates share another critical aspect of human-like mentalizing: what it means for someone to be ignorant or *unaware*. Understanding when others don't know something is an important aspect of human theory of mind. People successfully track other agents' ignorance across many contexts (e.g., when keeping or divulging secrets, when deciding what information to teach, when making communicative utterances intended to provide information, etc.). But do primates share this ability to reason about when others are unaware?

At first glance, the work reviewed previously might seem to show evidence that primates can track others' unawareness. For example, rhesus monkeys avoid stealing food from a person who is aware of their actions, but selectively take food from a person who is unaware (Flombaum & Santos, 2005, see also Hare et al., 2001 for similar evidence in chimpanzees). Similarly, macaques expect an aware agent to search for an object where she last saw it, but show no prediction about where an unaware person will search (Marticorena et al., 2011). These results have historically led researchers to argue that primates have an understanding of what it means for an agent to be unaware (Call & Santos, 2012; Call & Tomasello, 2008; Rosati et al., 2010; Whiten, 2013). Nevertheless, we and colleagues have recently proposed a different account of this pattern of performance (Horschler et al., 2019; Martin & Santos, 2016). While available results clearly suggest that primates treat aware agents *differently* from unaware agents, there is no clear evidence that primates make clear positive predictions about what an unaware agent should do. That is, when primates are asked to reason about an unaware agent, they often *no prediction* about what the agent will do (Marticorena et al., 2011) or simply react in a way consistent with their usual baseline behavior (e.g., taking food they want: Flombaum & Santos, 2005; taking the highest-value of two food items: Kaminski et al., 2008). Indeed, to our knowledge, there is only one published study in which primates are required to make a specific positive prediction about how an unaware experimenter should behave. In this study (Karg et al., 2015), chimpanzees played a foraging game with one of two experimenters: a cooperative experimenter who shared food and a competitive experimenter who stole food. Chimpanzees sat in front of an array of food rewards that were inaccessible to them but could be covered with an opaque screen before the other experimenter arrived. The logic was that chimpanzees should want to make the food rewards as visible to the cooperative experimenter as possible: food rewards that were not already covered should remain that way, and any covered food rewards should be revealed. In contrast, chimpanzees should show different performance when playing with the competitive experimenter; in this case, they should leave any covered food rewards hidden and cover up any visible food rewards to make the competitor ignorant. Karg and colleagues found that chimpanzees successfully opened the covered food rewards when playing with the cooperator (i.e., they knew that they could reveal foods to make the cooperator *aware*) but they failed to cover up the visible food rewards when playing with the competitor (i.e., they didn't seem to realize that they needed to make the competitor *unaware*). Apes therefore don't seem to realize what it means to make someone unaware. Yet again, when faced with an unaware experimenter, they default to the most obvious behavior (in this case, not moving any of the covers). We have argued that results like these suggest that in addition to lacking representations of others' beliefs, primates may also lack representations of others' unawareness (see Martin & Santos, 2016). That is, primates may not show a human-like understanding of others' belief or others' ignorance, despite their success in understanding others' awareness.

A Unified Theory for Nonhuman Theory of Mind

Given the complicated pattern of findings observed above, what can we conclude currently about the nature of primate mentalizing capacities? First, we argue that the results to date suggest clear evidence that primates can represent others' awareness. Many primate species successfully make predictions about what other agents see and know in a variety of different tasks. Moreover, primates use information about what agents are aware of to succeed in naturalistic competitive tasks. In this way, primates seem to possess one of the important aspects of human theory of mind capacities: the capacity to represent what others see and know (see also Martin & Santos, 2016 for a review of this awareness relations account).

In contrast, there is rather limited evidence that primates share our understanding of others' beliefs. Despite decades of experimental attempts, there is only one published report demonstrating that apes can succeed in a standard false belief test (Krupenye et al., 2016). This study also used an anticipatory looking measure that has recently generated some controversy in the developmental literature concerning whether human infants robustly show false belief reasoning on this task (e.g., Baillargeon et al., 2018). While the jury is still out about how to interpret these new ape anticipatory looking successes, the preponderance of false belief failures observed in comparative studies suggests that either primates cannot represent others' beliefs at all or that such representations are incredibly fragile and task specific. Taken together, then, there appears to be one area in which adult primates mentalize quite differently from adult humans: they seem (mostly) unable to represent other individuals' beliefs.

Finally, there is growing evidence of a second domain in which primate mentalizing may differ from that of humans: primates may lack the capacity to track unawareness in others. To date, most primate studies of unawareness show that primates make no positive predictions when faced with ignorant agents; they show no prediction when an agent who lacks awareness searches for a hidden object (Marticorena et al., 2011) and switch to default behaviors whenever a competitor is ignorant (Flombaum & Santos, 2005; Kaminski et al., 2008). The one study in which chimpanzees had the opportunity to actively make a competitor ignorant (Karg et al., 2015) found that primates fail to do so. Overall, primates' emerging pattern of performance suggests that they may represent others' awareness, but not others' unawareness (see also Martin & Santos, 2016 for a review).

What Is Missing?: Future Directions

Given the current evidence for primates' successes and failures in mentalizing, there are a few obvious next steps both for understanding the representations that primates use to make sense of other agents and for determining which aspects of human mentalizing are unique. First, more work is needed to better clarify if and when (some) primates do indeed track others' beliefs. As noted above, primates have long shown a consistent pattern of failures on false belief tasks (Call & Tomasello, 1999; Kaminski et al., 2008; Krachun et al., 2009; Marticorena et al., 2011; Martin & Santos, 2014; O'Connell & Dunbar, 2003), but one new study has argued that apes may be able to represent others' beliefs at least under very specific conditions (Krupenye et al., 2016). At the present time, it's not clear how to rectify these new findings with previous failures and thus future work could profit from replicating chimpanzees' performance on anticipatory looking tasks as well as testing other primates on related tasks. Moreover, researchers must think more about why some looking methods (e.g., anticipatory looking) are more likely to demonstrate successful belief reasoning than others (e.g., violation of expectation, as in Marticoriena et al., 2011; competitive tasks, as in Kaminski et al., 2008) where primates have previously shown task successes in different mentalizing abilities.

A second avenue of future research is to further test whether primates successfully make positive predictions about unaware agents. To date, few comparative researchers have designed studies of unawareness representations that require primates positive predictions about how an unaware agent will be have (see Karg et al., 2015 for an exception). Such studies would help us determine what (if anything) primates understand about others' ignorance. It's also worth noting that relatively little is known about human infants' unawareness representations; this too is a ripe area for future study, as we know little about infants' understanding of ignorance interacts with their early belief representations.

Third, and perhaps most importantly, future research must aim to better understand the nature of the representations that primates *do* possess: an understanding of others' states of awareness. For example, how and under what conditions do primates decide that a person is aware and turn on these awareness representations? And what kinds of situations cause primates to stop tracking an agent's awareness? Recent work has begun focusing on this latter question. In one recent study, Horschler et al. (2019) used an expectancy violation paradigm to test whether they could "break" a subject monkey's representation that an agent was aware of the location of an object. In the task, subjects saw a human agent watch a piece of fruit hidden in one of two boxes. After the agent's view was occluded, the fruit quickly moved outside of the box and back in. The agent then reached into one of the two boxes. If monkeys' representations of the agent's awareness are robust to irrelevant changes (e.g., a small irrelevant motion of the fruit), then subjects should look longer when the experimenter searches in the wrong location for the object. In contrast, if monkeys' representations are fragile enough to be disrupted by simple spatial manipulation of the object, then subjects should show no expectation about where the agent will search. Horschler and colleagues found this latter pattern of looking, suggesting that primates' awareness representations can be disrupted even by a quick movement of the target object while the person is looking away. Importantly, not all changes outside of a person's awareness seem to affect primates' awareness representations. Horschler and colleagues added a control in which a change irrelevant to the location of the object happened when the person wasn't paying attention (e.g., the box covering the fruit flipped open and immediately closed but the object remained stationary). In this case, monkeys were able to make a positive prediction about where the agent would search for the object, looking longer when the agent looked in the wrong (i.e., empty) box. These findings suggest that monkeys stop representing an agent as aware of an object's location when the object makes an irrelevant movement but not when there are changes to the target environment. In this regard, it seems that primates' awareness representations may be more nuanced than previously thought. Moving forward, future work should critically test whether other state changes of the target object (e.g., location, physical appearance) are enough to break these awareness representations. A better understanding of these mechanisms will thus be important not only for a full account of primate mentalizing but also for gaining more clarity on which aspects of human mentalizing are unique.

There is also a need to better explore how primates develop these awareness representations in the first place. Although much is known about the early development of human theory of mind representations (e.g., Helming et al., 2014; Wellman & Liu, 2004; Scott & Baillargeon, 2017; Sodian, 2011), to date there are few studies exploring the ontogenetic origins of primate mentalizing. This is unfortunate, as understanding the developmental history of primates' theory of mind representations would provide critical hints as to whether they emerge along the same timeline as early human mentalizing capacities (see review in Rosati, Wobber, Hughes, & Santos, 2014).

Indeed, recent empirical work hints that some primates may show a different developmental pattern in their early mentalizing than humans do; specifically, one species of primates (the rhesus macaque) appears to develop awareness representations in a delayed time course (Arre, Clark, & Santos, 2020; Rosati, Arre, Platt, &

Santos, 2016) relative to humans. This new work suggests that some primates may require more experience interacting with other social agents in order to develop sophisticated awareness representations. Moving forward, the field of comparative cognitive development may provide critical tests for the types of experiences required for different mentalizing capacities to come online.

Finally, having established that primates share some aspects of human-like mentalizing capacities, it is important to explore the neural basis of these socio-cognitive abilities. While human social cognitive neuroscience has made great strides in understanding the mechanisms by which humans mentalize about their future selves and others (Leshinskaya, Contreras, Caramazza, & Mitchell, 2017; Macrae et al., 2017; Tamir, Thornton, Contreras, & Mitchell, 2016; Thornton & Mitchell, 2017), very little is known about the neural mechanisms primates use to track what others know and perceive. Although it is often hypothesized that there is a high level of conservation between human and nonhuman primate functional neuroanatomy (Chang et al., 2013, 2015; Ghazanfar & Santos, 2004; Platt, Seyfarth, & Cheney, 2016; Platt & Spelke, 2009), little research has investigated whether similar neural mechanisms underlie mentalizing in our primate relatives. Excitingly, comparative researchers have now developed a number of experimental methodologies that can be adapted into neurophysiological preparations. In addition, we now know much more about the mentalizing abilities of those primate species specifically used in neuroscientific investigations (e.g., rhesus macaques, for a review see Drayton & Santos, 2016). In this way, the empirical stage is now set to begin exploring the neural mechanisms underlying theory of mind capacities in primates.

The past four decades have revealed much about the nature of primate' understanding of other agents. Our continued hope is that a better understanding of primates' successes and failures on mentalizing tasks can help cognitive and neural scientists to better understand not just how primates think about other minds but what primates mentalizing reveals about the cognitive and neural basis of our own species' mentalizing as well.

References

Amici, F., Aureli, F., Visalberghi, E., & Call, J. (2009). Spider monkeys (Ateles geoffroyi) and capuchin monkeys (Cebus apella) follow gaze around barriers: Evidence for perspective taking. *Journal of Comparative Psychology, 123*(4), 368.

Arre, A. M., Clark, C. S., & Santos, L. R. (2020). Do young rhesus monkeys know what others see?: A comparative developmental perspective. *American Journal of Primatology*, e23054.

Arre, A.M., Stumph, E., & Santos, L.R. (2021). Macaque species with varying social tolerance show no differences inunderstanding what other agents perceive. *Animal Cognition*.

Astington, J. W., & Gopnik, A. (1991). Theoretical explanations of children's understanding of the mind. *British Journal of Developmental Psychology, 9*(1), 7–31.

Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Invited commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development, 46*, 112–124.

Bettle, R., & Rosati, A. G. (2019). Flexible gaze-following in rhesus monkeys. *Animal Cognition, 22*, 673–686.

Botting, J. L., Wiper, M. L., & Anderson, J. R. (2011). Brown (Eulemur fulvus) and ring-tailed lemurs (Lemur catta) use human head orientation as a cue to gaze direction in a food choice task. *Folia Primatologica, 82*(3), 165–176.

Bräuer, J., Call, J., & Tomasello, M. (2005). All great ape species follow gaze to distant locations and around barriers. *Journal of Comparative Psychology, 119*(2), 145.

Bräuer, J., Kaminski, J., Riedel, J., Call, J., & Tomasello, M. (2006). Making inferences about the location of hidden food:social dog, causal ape. *Journal of Comparative Psychology, 120*(1), 38.

Burkart, J., & Heschl, A. (2006). Geometrical gaze following in common marmosets (Callithrix jacchus). *Journal of Comparative Psychology, 120*(2), 120.

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition, 112*(2), 337–342.

Buttelmann, D., Buttelmann, F., Carpenter, M., Call, J., & Tomasello, M. (2017). Great apes distinguish true from false beliefsin an interactive helping task. *PLoS One, 12*(4), e0173793.

Byrne, R. W., & Bates, L. A. (2010). Primate social cognition: Uniquely primate, uniquely social, or just unique? *Neuron, 65*(6), 815–830.

Call, J., & Santos, L. R. (2012). Understanding other minds. In *The evolution of primate societies* (pp. 664–681). Chicago, IL: University of Chicago Press.

Call, J., & Tomasello, M. (1999). A nonverbal false belief task: The performance of children and great apes. *Child Development, 70*(2), 381–395.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences, 12*(5), 187–192.

Chang, S. W., Brent, L. J., Adams, G. K., Klein, J. T., Pearson, J. M., Watson, K. K., & Platt, M. L. (2013). Neuroethology of primate social behavior. *Proceedings of the National Academy of Sciences, 110*(Suppl 2), 10387–10394.

Chang, S. W., Fagan, N. A., Toda, K., Utevsky, A. V., Pearson, J. M., & Platt, M. L. (2015). Neural mechanisms of social decision-making in the primate amygdala. *Proceedings of the National Academy of Sciences, 112*(52), 16012–16017.

Crockford, C., Wittig, R. M., Mundry, R., & Zuberbühler, K. (2012). Wild chimpanzees inform ignorant group members of danger. *Current Biology, 22*(2), 142–146.

Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences, 1*(4), 568–570.

Drayton, L. A., & Santos, L. R. (2016). A decade of theory of mind research on Cayo Santiago: Insights into rhesus macaque social cognition. *American Journal of Primatology, 78*(1), 106–116.

Drayton, L. A., & Santos, L. R. (2017). Do rhesus macaques, Macaca mulatta, understand what others know when gaze following? *Animal Behaviour, 134*, 193–199.

Drayton, L. A., & Santos, L. R. (2018). What do monkeys know about others' knowledge? *Cognition, 170*, 201–208.

Emery, N. J., Lorincz, E. N., Perrett, D. I., Oram, M. W., & Baker, C. I. (1997). Gaze following and joint attention in rhesus monkeys (Macaca mulatta). *Journal of Comparative Psychology, 111*(3), 286.

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology, 50*(1), 21–45.

Flombaum, J. I., & Santos, L. R. (2005). Rhesus monkeys attribute perceptions to others. *Current Biology, 15*(5), 447–452.

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review, 5*(3), 490–495.

Ghazanfar, A. A., & Santos, L. R. (2004). Primate brains in the wild: The sensory bases for social interactions. *Nature Reviews Neuroscience, 5*(8), 603.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language, 7*(1–2), 145–171.

Hains, S. M., & Muir, D. W. (1996). Infant sensitivity to adult eye direction. *Child Development, 67*(5), 1940–1951.

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour, 59*(4), 771–785.

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour, 61*(1), 139–151.

Hare, B., & Tomasello, M. (2004). Chimpanzees are more skilful in competitive than in cooperative cognitive tasks. *Animal Behaviour, 68*(3), 571–581.

Hare, B., Call, J., & Tomasello, M. (2006). Chimpanzees deceive a human competitor by hiding. *Cognition, 101*(3), 495–514.

Hattori, Y., Kano, F., & Tomonaga, M. (2010). Differential sensitivity to conspecific and allospecific cues in chimpanzees and humans: A comparative eye-tracking study. *Biology Letters, 6*(5), 610–613.

Helming, K. A., Strickland, B., & Jacob, P. (2014). Making sense of early false-belief understanding. *Trends in Cognitive Sciences, 18*(4), 167–170.

Hirata, S., & Matsuzawa, T. (2001). Tactics to obtain a hidden food item in chimpanzee pairs (Pan troglodytes). *Animal Cognition, 4*(3-4), 285–295.

Hood, B. M., Willen, J. D., & Driver, J. (1998). Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science, 9*(2), 131–134.

Horschler, D. J., Santos, L. R., & MacLean, E. L. (2019). Do non-human primates really represent others' ignorance? A test of the awareness relations hypothesis. *Cognition, 190*, 72.

Hostetter, A. B., Russell, J. L., Freeman, H., & Hopkins, W. D. (2007). Now you see me, now you don't: Evidence that chimpanzees understand the role of the eyes in attention. *Animal Cognition, 10*(1), 55.

Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition, 109*(2), 224–234.

Karg, K., Schmelz, M., Call, J., & Tomasello, M. (2015). The goggles experiment: Can chimpanzees use self-experience to infer what a competitor can see? *Animal Behaviour, 105*, 211–221.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science, 330*, 1830–1834.

Krachun, C., Carpenter, M., Call, J., & Tomasello, M. (2009). A competitive nonverbal false belief task for children and apes. *Developmental Science, 12*(4), 521–535.

Krupenye, C. (2020). Theory of mind in great apes and its evolutionary origins. In K. Ochsner & M. Gilead (Eds.), *The neural bases of mentalizing*. New York, NY: Springer.

Krupenye, C., & Call, J. (2019). Theory of mind in animals: Current and future directions. *Wiley Interdisciplinary Reviews: Cognitive Science, 10*, e1503.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science, 354*(6308), 110–114.

Leshinskaya, A., Contreras, J. M., Caramazza, A., & Mitchell, J. P. (2017). Neural representations of belief concepts: A representational similarity approach to social semantics. *Cerebral Cortex, 27*(1), 344–357.

Lyons, D. E., & Santos, L. R. (2006). Ecology, domain specificity, and the origins of theory of mind: is competition thecatalyst?. *Philosophy Compass, 1*(5), 481–492.

Luo, Y. (2011). Do 10-month-old infants understand others' false beliefs? *Cognition, 121*(3), 289–298.

Luo, Y., & Johnson, S. C. (2009). Recognizing the role of perception in action at 6 months. *Developmental Science, 12*(1), 142–149.

MacLean, E. L., & Hare, B. (2012). Bonobos and chimpanzees infer the target of another's attention. *Animal Behaviour, 83*(2), 345–353.

Macrae, C. N., Mitchell, J. P., Golubickis, M., Ho, N. S., Sherlock, R., Parlongo, R., … Christian, B. M. (2017). Saving for your future self: The role of imaginary experiences. *Self and Identity, 16*(4), 384–398.

Marticorena, D. C., Ruiz, A. M., Mukerji, C., Goddu, A., & Santos, L. R. (2011). Monkeys represent others' knowledge but nottheir beliefs. *Developmental Science, 14*(6), 1406–1416.

Martin, A., & Santos, L. R. (2014). The origins of belief representation: Monkeys fail to automatically represent others' beliefs. *Cognition, 130*(3), 300–308.

Martin, A., & Santos, L. R. (2016). What cognitive representations support primate theory of mind?. *Trends in CognitiveSciences, 20*(5), 375–382.

Melis, A. P., Call, J., & Tomasello, M. (2006). Chimpanzees (Pan troglodytes) conceal visual and auditory information from others. *Journal of Comparative Psychology, 120*(2), 154.

Neiworth, J. J., Burman, M. A., Basile, B. M., & Lickteig, M. T. (2002). Use of experimenter-given cues in visual co-orienting and in an object-choice task by a New World monkey species, Cotton Top Tamarins (Saguinus oedipus). *Journal of Comparative Psychology, 116*(1), 3.

O'Connell, S., & Dunbar, R. I. M. (2003). A test for comprehension of false belief in chimpanzees. *Evolution and Cognition, 9*(2), 131–140.

Okamoto-Barth, S., Call, J., & Tomasello, M. (2007). Great apes' understanding of other individuals' line of sight. *Psychological Science, 18*(5), 462–468.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255–258.

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 362*(1480), 731–744.

Platt, M. L., Seyfarth, R. M., & Cheney, D. L. (2016). Adaptations for social cognition in the primate brain. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 371*(1687), 20150096.

Platt, M. L., & Spelke, E. S. (2009). What can developmental and comparative cognitive neuroscience tell us about the adult human brain? *Current Opinion in Neurobiology, 19*(1), 1.

Poulin-Dubois, D., Rakoczy, H., Burnside, K., Crivello, C., Dörrenberg, S., Edwards, K., … Perner, J. (2018). Do infants understand false beliefs? We don't know yet–A commentary on Baillargeon, Buttelmann and Southgate's commentary. *Cognitive Development, 48*, 302–315.

Povinelli, D. J., & Eddy, T. J. (1996). Chimpanzees: Joint visual attention. *Psychological Science, 7*(3), 129–135.

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development, 46*, 40–50.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(4), 515–526.

Pylyshyn, Z. W. (1978). When is attribution of beliefs justified?. *Behavioral and Brain Sciences, 1*(4), 592–593.

Rosati, A. G., Arre, A. M., Platt, M. L., & Santos, L. R. (2016). Rhesus monkeys show human-like changes in gaze following across the lifespan. *Proceedings of the Royal Society B: Biological Sciences, 283*(1830), 20160376.

Rosati, A. G., & Hare, B. (2009). Looking past the model species: Diversity in gaze-following skills across primates. *Current Opinion in Neurobiology, 19*(1), 45–51.

Rosati, A. G., Santos, L. R., & Hare, B. (2010). Primate social cognition: Thirty years after Premack and Woodruff. *Primate Neuroethology, 1*(9), 117–144.

Rosati, A. G., Wobber, V., Hughes, K., & Santos, L. R. (2014). Comparative developmental psychology: How is human cognitive development unique? *Evolutionary Psychology, 12*(2), 448–473.

Rubio-Fernández, P., & Geurts, B. (2013). How to pass the false-belief task before your fourth birthday. *Psychological Science, 24*(1), 27–33.

Ruiz, A., Gómez, J. C., Roeder, J. J., & Byrne, R. W. (2009). Gaze following and gaze priming in lemurs. *Animal Cognition, 12*(3), 427–434.

Sandel, A. A., MacLean, E. L., & Hare, B. (2011). Evidence from four lemur species that ringtailed lemur social cognition converges with that of haplorhine primates. *Animal Behaviour, 81*(5), 925–931.

Santos, L. R., Nissen, A. G., & Ferrugia, J. A. (2006). Rhesus monkeys, Macaca mulatta, know what others can and cannot hear. *Animal Behaviour, 71*(5), 1175–1181.

Schmelz, M., Call, J., & Tomasello, M. (2011). Chimpanzees know that others make inferences. *Proceedings of the National Academy of Sciences, 108*(7), 3077–3079.

Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences, 21*(4), 237–249.

Seed, A., & Tomasello, M. (2010). Primate cognition. *Topics in Cognitive Science, 2*(3), 407–419.

Shepherd, S. V., & Platt, M. L. (2008). Spontaneous social orienting and gaze following in ringtailed lemurs (Lemur catta). *Animal Cognition, 11*(1), 13.

Sodian, B. (2011). Theory of mind in infancy. *Child Development Perspectives, 5*(1), 39–43.

Song, H. J., & Baillargeon, R. (2007). Can 9.5-month-old infants attribute to an agent a disposition to perform a particular action on objects? *Acta Psychologica, 124*(1), 79–105.

Song, H. J., Baillargeon, R., & Fisher, C. (2005). Can infants attribute to an agent a disposition to perform a particular action? *Cognition, 98*(2), B45–B55.

Symons, L. A., Hains, S. M., & Muir, D. W. (1998). Look at me: Five-month-old infants' sensitivity to very small deviations in eye-gaze during social interactions. *Infant Behavior and Development, 21*(3), 531–536.

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences, 113*(1), 194–199.

Thornton, M. A., & Mitchell, J. P. (2017). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex, 28*(10), 3505–3520.

Tomasello, M. (2000). Culture and cognitive development. *Current Directions in Psychological Science, 9*(2), 37–40.

Tomasello, M. (2018). Great apes and human development: A personal history. *Child Development Perspectives, 12*(3), 189–193.

Tomasello, M. (2019). Evolutionary foundations. In *Becoming human: A theory of ontogeny* (pp. 10–44). Cambridge, MA: Belknap Press.

Tomasello, M., Call, J., & Hare, B. (1998). Five primate species follow the visual gaze of conspecifics. *Animal Behaviour, 55*(4), 1063–1069.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 675–691.

Tomasello, M., Hare, B., & Agnetta, B. (1999). Chimpanzees, Pan troglodytes, follow gaze direction geometrically. *Animal Behaviour, 58*(4), 769–777.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in youngchildren's understanding of deception. *Cognition, 13*(1), 103–128.

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everydaypsychology. *Cognition, 35*(3), 245–275.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655–684.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child development, 75*(2), 523–541.

Wellman, H. M. (2011). *Developing a theory of mind. The Wiley-Blackwell handbook of childhood cognitive development* (2nd ed., pp. 268–284). New York, NY: Wiley-Blackwell.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75*(2), 523–541.

Whiten, A. (2013). Humans are not alone in computing how others see the world. *Animal Behaviour, 86*(2), 213–221.

Empathic Accuracy: Empirical Overview and Clinical Applications

Céline Hinnekens, William Ickes, Liesbet Berlamont, and Lesley Verhofstadt

Introduction

This chapter tries to present a comprehensive picture of the ongoing literature on empathic accuracy. First, we describe the conceptualization and measurement of empathic accuracy. Second, we provide a brief overview of research on the predictors of empathic accuracy. Third, we examine the role of empathic accuracy in relationships, as revealed by relevant theory and research. Finally, we examine how empathic accuracy is related to psychotherapy, to various clinical disorders, and to other clinically relevant outcomes.

Conceptualization

Empathic accuracy was first identified as an important phenomenon within clinical and consulting psychology, but has been studied within many disciplines over the past three decades. Rogers (1957) defined the term *accurate empathy* as the therapist's ability to accurately discern the contents of the client's thoughts and feelings as they change over time. He referred to a complex process that is motivated by the therapist's desire to observe and understand the inner perceptual world of the client.

C. Hinnekens (✉) · L. Berlamont · L. Verhofstadt
Department of Experimental-Clinical and Health Psychology, Ghent University,
Ghent, Belgium
e-mail: celine.hinnekens@ugent.be; Liesbet.berlamont@ugent.be;
Lesley.verhofstadt@ugent.be

W. Ickes
Department of Psychology, University of Texas at Arlington, Arlington, TX, USA
e-mail: ickes@uta.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_8

It required not only attending to the client's explicit communication, but also tracking the client's continuous flow of internal cognitive and emotional interpretations.

Roger's concept of *accurate empathy* was the direct precursor for Ickes's concept of empathic accuracy, which is defined as "the extent to which perceivers understand a target's episodic thoughts and feelings as they appear spontaneously during the course of a natural interaction" (Ickes, 1993, p. 588). The main difference between the two concepts is that *accurate empathy* focuses on the empathic process itself, whereas *empathic accuracy* focuses on the degree of accuracy the perceiver achieves as an outcome of this process.

It should be emphasized that empathic accuracy is conceptualized as the outcome of an interpersonal, multidimensional (i.e., influenced by many underlying predictors), and situation-specific process. This process is affected by characteristics of the target (the person who is experiencing the thoughts and feelings), the perceiver (the person who infers the target's thoughts and feelings), their past and present relationship, and the specific situation in which they find themselves. Adding to this complexity, empathic accuracy can result from dynamically unfolding dyadic process in which two individuals both assume the dual roles of target and perceiver while interacting with each other, or it can occur in a context in which two interaction partners (targets) are observed by a third party (perceiver) who is not participating in their interaction.

In either case, the perceiver attempts to achieve an accurate assessment of the inner experiences of the target person by drawing on a variety of cognitive and emotional resources that include observation, memory, reasoning, inference, analogy-to-self (projection), and emotional contagion (Ickes, 1997). Therefore, the perceiver must detect and evaluate the available informational cues provided by the target, and then integrate and interpret them with respect to both the situation-specific context of the current interaction and their schemas about the individual targets, the nature of their relationship, and their previous interactions (if the interaction partners have a shared history).

Operationalization

Before 1990, research on empathic accuracy was mainly situated in the context of training programs for students in clinical psychology who were learning to conduct psychotherapy. The studies conducted during this period examined how accurately students inferred clients' thoughts and feelings during either real or simulated psychotherapy sessions (Ickes, 2003). Although these studies were useful as training exercises, their scientific value was limited by the fact that the inferences made by the student trainee (perceiver) were typically not compared with actual thoughts and feelings reported by the client (target), but rather with inferences that were made by the supervising psychotherapist (Ickes, 2003). Therefore, the resulting "accuracy scores" were more of a measure of faith in the expertise of the supervisor than an objective measure of clinical mind-reading performance.

Taking a large step toward greater objectivity, the clinical psychologist Nathan Kagan (1977) introduced the *standard stimulus paradigm* (SS-paradigm) in which the participants watched a standard set of videos that depicted the unstructured interactions of client-therapist pairs. The participants were asked to observe these videos, which were paused at times when the client had reported a specific thought or feeling. At each of these times, the participants were required to select the client's actual reported thought or feeling from a set of plausible multiple-choice alternatives. In this paradigm, the perceiver's empathic accuracy was calculated quite simply as the proportion of correct choices made. Note, however, that the Kagan paradigm did not require the perceiver to generate the client's actual thought or feeling, but merely to select it from a set of previously prepared alternatives.

To address this limitation, the social psychologist William Ickes and his colleagues developed a way to measure empathic accuracy within the context of their previously developed dyadic interaction paradigm (DI-paradigm; Ickes, 1982; Ickes, Stinson, Bissonnette, & Garcia, 1990; Ickes & Tooke, 1988). This paradigm allows empathic accuracy to be measured in a more naturalistic setting, and in an objective and reliable manner. The original DI-paradigm (also known as the *standard empathic accuracy assessment procedure*) includes two interacting partners who are both a perceiver and a target. During the first stage of the procedure, the two participants spontaneously interact with each other in a room where they are alone but are secretly being videotaped. During the second phase, both participants—working independently—observe a video recording of their just-completed interaction and use it to make a record of the specific thoughts and feelings they had during the interaction and exactly when they occurred. During the third phase, they watch the video recording a second time while the video is paused at the particular moments when the target partner reported having had a thought or feeling, and are asked to infer the partner's specific thought or feeling at each of these points.

After all data are collected, the similarity between each actual and inferred thought or feeling is rated by independent coders, using a 3-point (0–2) scale (i.e., 0 = different content from the actual thought or feeling, 1 = similar, but not the same, content as the actual thought or feeling, and 2 = essentially the same content as the actual thought or feeling). The accuracy points earned by the perceiver divided by the total accuracy points possible results in an empathic accuracy score between 0 and 100. Variations on the paradigm mainly concern the type of relationship between the interaction partners (e.g., strangers, friends, dating, or marriage partners; Thomas & Fletcher, 2003) or the conversation type (e.g., spontaneous conversation, therapy session, support and conflict interactions; Devoldre, Davis, Verhofstadt, & Buysse, 2010; Hinnekens, Loeys, De Schryver, & Verhofstadt, 2018; Ickes et al., 1990; Marangoni, Garcia, Ickes, & Teng, 1995).

The DI-paradigm, in which each dyad member is both a perceiver and a target person, differs in one very important respect from the SS-paradigm in which the participants take only the role of perceiver. In the SS-paradigm, all of the perceivers observe the same targets, which allows researchers to compare the accuracy scores of the different perceivers for the same target or the different perceivers' average accuracy scores over various targets. Such comparisons are very difficult to make in the DI-paradigm.

Predictors of Empathic Accuracy

In everyday life, it is often said that some people are better than others at the task of inferring target persons' thoughts and feelings. These comparisons suggest that empathic accuracy is the product of either an inherited ability level or one that develops over time. Either way, empirical evidence for this assumption of relatively stable individual differences should be found in the possibility to generate a list of reliable personality predictors of a "good" perceiver. Below, we present a brief review of the research findings pertaining to perceiver characteristics, followed by empirical evidence for the additional importance of target characteristics in predicting empathic accuracy (see Hodges, Lewis, & Ickes, 2015 for an overview).

Perceiver Characteristics

Interpersonal Sensitivity

Although one would expect that perceivers who score high on measures of interpersonal sensitivity would also achieve higher levels of empathic accuracy, the relevant findings are generally mixed. For example, one study reported no significant associations between empathic accuracy, as measured using the DI-paradigm, and the same perceivers' performance on either the Diagnostic Analysis of Nonverbal Accuracy (DANVA; Nowicki & Duke, 1994)—in which perceivers have to assign emotional labels to facial expressions—or on the Interpersonal Perception Task (IPT; Costanzo & Archer, 1989)—in which perceivers must evaluate interactions based on (non-)verbal cues (Lewis & Hodges, 2009).

The Interpersonal Reactivity Index (IRI; Davis, 1983), one of the most frequently used self-report questionnaires measuring components of dispositional empathy, has also proved to be inconsistently correlated with empathic accuracy. Specifically, the IRI-subscales that are conceptually most in line with the concept of empathic accuracy, namely perspective-taking and empathic concern, were either uncorrelated or only weakly and ambiguously correlated with empathic accuracy in an early DI-paradigm study (Stinson & Ickes, 1992). However, scores on the Balanced Emotional Empathy Scale (BEES; Mehrabian & Epstein, 1972), which measures the degree of empathy for emotions, were associated with empathic accuracy in a different study using the SS-paradigm, one in which the target persons showed a relatively high degree of expressiveness (Zaki, Bolger, & Ochsner, 2008).

In summary, individuals' self-reports of interpersonal sensitivity are neither strongly nor consistently related to their levels of empathic accuracy. One possible explanation for these findings is that most people are poor at assessing their own ability to accurately infer other people's thoughts and feelings.

IQ, Academic Performance, and Verbal Intelligence

The investigation of the link between intelligence quotient (IQ) and empathic accuracy also yielded mixed and somewhat qualified results. In two studies, IQ and academic performance were found to be significant predictors of college students' empathic accuracy (Ickes et al., 1990; Ponnet, Buysse, Roeyers, & De Clercq, 2008), but other studies reported nonsignificant or very weak associations (e.g., Ponnet, Roeyers, Buysse, De Clercq, & Van der Heyden, 2004). A study by Ickes, Buysse, et al. (2000) found that verbal intelligence was a potential predictor of empathic accuracy, but only in men.

Sex and Gender

In line with the assumption that certain people are more empathically accurate than others, a common gender-based stereotype asserts that women are more empathic than men. A relevant review article (Hodges, Laurent, & Lewis, 2011) concluded that although some studies have found significant gender differences in empathic accuracy that favor female perceivers, many studies have failed to find such a difference. Interestingly, women were consistently more accurate than men only in situations in which the gender-based stereotype was implicitly or explicitly evoked. This pattern suggests that women—but not men—put more effort into making accurate inferences as a way of trying to meet the expectations of the female gender role (Ickes, Gesn, & Graham, 2000). Women do appear to be better at inferring the emotional significance of nonverbal cues (Hall, 1984), but this is only one component of empathic accuracy—and a relatively minor one at that (see Gesn & Ickes, 1999; Hall & Schmid Mast, 2007). As Ickes, Gesn, et al. (2000) stated, a significant gender difference in empathic accuracy is often not found, and when it is found it appears to be primarily motivational in nature.

Conclusion

In summary, the search for *perceiver characteristics* predicting empathic accuracy has yielded mixed and even contradictory results. It is noteworthy that most studies that have found significant results have measured empathic accuracy using the SS-paradigm (the one in which the performance of different perceivers can be better compared, because they are all inferring the same set(s) of thought/feeling inferences). On the other hand, in studies using the DI-paradigm (in which different perceivers infer different sets of thoughts and feelings), few or no perceiver characteristics have been found to be significant.

Target Characteristics

The inconclusive evidence for reliable *perceiver characteristics* as predictors of empathic accuracy led researchers to turn their attention to potential *target characteristics*. This research focused on individual differences between targets that might account for how easy or difficult it is for a perceiver to infer the thoughts and feelings of that particular target.

Readability

This concept refers to how "readable" or transparent the target's thoughts and feelings are in comparison to those of other targets. Several studies have operationalized readability as a global index of how difficult it is to infer a target's thoughts and feelings (e.g., the inferential difficulty index proposed by Marangoni et al., 1995). This index is computed as the average judgment by a set of trained raters of how easy or difficult it would be to infer each of the actual thoughts and feelings based on the verbal and nonverbal cues that were available right before it occurred. As expected, the readability index has been found to be significantly correlated with empathic accuracy, suggesting that some targets are less/more transparent and "readable" than others. Readability had an impact on empathic accuracy that was most significant for interaction partners who did not know each other in advance and thus could only rely on each other's immediate verbal and nonverbal cues as a basis from which to make empathic inferences (because they had no prior information or knowledge about each other).

Sex and Gender

An interesting but rarely examined question concerns the existence of gender differences in the readability of targets—that is, are female targets more transparent and "readable" than male targets? Hall (1984) found that women emit more obvious nonverbal cues than men, but, in contradiction to this finding, a study by Simpson et al. (2011) found that men were easier to read than women are. A speculative explanation for these contradictory findings is that men might compensate for their lack of expressiveness and self-disclosure by showing greater consistency between their verbal cues (what they say and how they say it) and the content of the actual thoughts and feelings they are having.

Conclusion

The results of a meta-analytic study have revealed that there is far more target-variance than perceiver-variance in empathic accuracy scores, suggesting that certain features of the target affect the perceiver's empathic accuracy more than characteristics of the perceiver do (Ickes, Buysse, et al., 2000). Much remains to be learned, however, about *which* target characteristics are the most important ones, and why they are.

Motivational Influences on Empathic Accuracy

A different line of research has explored the effects of *motivational influences* on empathic accuracy. Two categories of motivational influences have been explored: motives that are evoked by the situation the perceiver is in and motives that derive from the perceiver's own personality.

Situationally Evoked Motives

The situationally evoked motives that have been explored in experimental research include money (being paid more for greater empathic accuracy), the prospect of success with attractive women, social recognition, and so on. The pattern of findings is, once again, mixed. In an early study using the DI-paradigm, Klein and Hodges (2001) found that a financial incentive led to greater accuracy—particularly for men. However, a subsequent study by Hall et al. (2009) revealed no effect of rewards, including financial incentives.

Another situationally induced motive concerns the attractiveness of the partner. In an early study of mixed-sex dyads, Ickes et al. (1990) found that perceivers were more accurate when their opposite-sex partners (new acquaintances) were physically attractive. Presumably, this effect occurred because the perceiver was more motivated to get to know an attractive target person, resulting in greater empathic accuracy (Ickes et al., 1990). Ickes and Hodges (2013) have speculated that other target variables, such as a challenging level of intelligence or a charismatic personality, might also have a positive influence on the perceiver's motivation to be accurate.

Personality-Based Motives

Some motives derive from the personality of the perceiver, with socially relevant motives being of particular interest. For example, Pickett, Gardner, and Knowles (2004) found that individuals who have a higher need to belong to a social group and to be connected with others measured with the Need to Belong Scale (Leary, Kelly, & Schreindorfer, 2001) performed better on an SS-paradigm task. In this case, the need to belong is assumed to motivate greater sensitivity and receptivity to the thoughts and feelings of others in order to form and maintain desired social relationships.

In other relevant research, anxiously attached women have been found to be more accurate at detecting the thoughts and feelings of their partners in relationship-threatening situations (e.g., when the partner is being interviewed by a physically attractive woman; Dugosh, 2001; more details on the influence of threat are described below). Conversely, securely attached women are generally less accurate when placed in threatening situations, possibly because an (unconscious) motive triggers inaccuracy for their partner's threatening thoughts and feelings in order to protect and maintain the relationship (Simpson, Oriña, & Ickes, 2003). Finally, people with an avoidant attachment style are found to be less accurate than others, independent of the situation (Simpson et al., 2011), expressing a general lack of interest in what their relationship partner is thinking or feeling.

Conclusion

These and other findings (see Smith, Ickes, Hall, & Hodges, 2011) lead to the important conclusion that empathic accuracy can either increase or decrease (be "dialed up" or "dialed down"), depending on the strength of the perceiver's current motive to be accurate or inaccurate. As we have seen, some of these motives are situationally evoked but others derive from the perceiver's own personality.

Empathic Accuracy in Relationships

Empirical research on the role of empathic accuracy in personal relationships began by investigating the importance of familiarity with an interaction partner. Later studies have led to a more nuanced and dynamic view of how empathic accuracy operates within our close relationships, including couple and family relationships.

Familiarity with the Other

When people interact with each other as strangers, they typically earn an average of only 17–22% of the available "accuracy points" when they attempt to infer each other's thoughts and feelings (Hinnekens et al., 2018; Ickes, 2011). The best mind-reading performances—the rare outliers—achieve about 55% of the available "accuracy points," which leaves nearly half of the potential performance range (45%) empty of cases. Still, previous research has revealed that, as expected, empathic accuracy varies as a function of the "type of relationship" between the target and the perceiver. In studies testing the so-called *acquaintanceship effect*, the empathic accuracy of strangers averages 20%, significantly lower than the average of 30% achieved by friends (Stinson & Ickes, 1992).

In light of these findings, it is surprising that the relationship between empathic accuracy and relationship duration in married couples has proved to be *negative* (Hinnekens, 2017; Kilpatrick, Bissonnette, & Rusbult, 2002; Thomas, Fletcher, & Lange, 1997). This finding suggests that, following the "honeymoon period," empathic accuracy declines as the partners settle in to their respective roles and focus less exclusively on each other. Over time, the partners appear to pay less attention to the actual cues that are relevant during their current interaction and instead infer each other's thoughts and feelings by relying on partner-specific schemas that reflect their historical knowledge of each other and their shared history as interaction partners (McLeod & Chaffee, 1973; Wegner, Giuliano, & Hertel, 1985). Although these schema-based inferences are somewhat less accurate, they are also more efficient—occurring automatically and resulting in empathic inferences that are typically "good enough" but not optimal.

Intimate Relationships

Partners must be at least relatively accurate ("good enough" or better) if they want to effectively coordinate their individual and shared actions and maintain a satisfying and stable relationship (Ickes & Hodges, 2013). Studies of empathic accuracy in intimate relationships reveal that partners are, at best, only moderately good at inferring each other's thoughts and feelings (Hinnekens, Ickes, De Schryver, & Verhofstadt, 2016; Thomas et al., 1997; Verhofstadt et al., 2016). According to Ickes (2011), empathic accuracy averages around 30–35% for married partners in research conducted in Texas. Research conducted in Belgium has found even lower empathic accuracy for married and cohabiting partners, averaging around 20% (Hinnekens et al., 2016; Verhofstadt et al., 2016; Verhofstadt, Buysse, Ickes, Davis, & Devoldre, 2008). And, not surprisingly, clinical observation has revealed that couples seeking therapy frequently complain about a lack of mutual understanding and "misreading" in their relationship (Gurman, 2008).

Couples' Interactional Behavior

In non-distressed relationships, situationally taking the perspective of one's partner is considered to be a necessary first step that enables partners to accommodate and adapt to each other on a daily basis. Research based on this assumption has therefore examined empathic accuracy within two key interactional domains in marriage: how partners help each other cope with personal stressors (i.e., support interactions) and how they deal with relationship stressors (i.e., conflict interactions).

In studies relevant to the first domain, the perceiver's empathic accuracy has been positively associated with the perceiver's ability to provide effective instrumental support to the partner (e.g., giving advice) and negatively associated with the perceiver's provision of negative types of support (e.g., criticizing the partner, minimizing the problem) (Devoldre et al., 2010; Verhofstadt et al., 2008, 2016). In studies relevant to the second domain, the perceiver's empathic accuracy also appears to play a constructive role in partners' conflict interactions, such that both partners exhibit more adequate problem-solving and accommodative behavior when they are more empathically accurate (Kilpatrick et al., 2002; Rusbult, Verette, Whitney, Slovik, & Lipkus, 1991). Taken together, these findings suggest that empathic accuracy fosters pro-relationship behavior when partners are faced with daily stressors that have their source either outside or inside the relationship (Bodenmann, 2005).

Relationship Satisfaction

Given the positive effects of empathic accuracy in marital support and marital conflict interactions, we would expect that empathic accuracy is also positively associated with partners' relationship satisfaction.

Indeed, in an article by Sillars and Scott (1983) that reviewed the early literature on understanding in couples—before the start of research on empathic accuracy and the DI-paradigm by Ickes et al. (1990)—the authors concluded that congruency between partners' perceptions (i.e., the presence of a shared perceptual reality) is central to relationship adjustment and satisfaction. Sillars and Scott cited a large number of studies that revealed positive associations between marital adjustment and the partners' understanding of each other's attitudes, expectations, and self-perceptions (e.g., Christensen & Wallace, 1976; Corsini, 1956; Dymond, 1954; Ferguson & Allen, 1978; Guthrie & Noller, 1988; Laing, Phillipson, & Lee, 1966; Luckey, 1960; Stuckert, 1963). Based on the widespread perception that more understanding is good for relationships, a dominant narrative of advice concerning couples' communication strategies emerged, one that emphasized the importance of self-disclosure to facilitate mutual understanding (Bochner, 1981).

Complementing the early findings that relate the congruence in partner's perceptions to their marital adjustment, a more recent meta-analysis involving 20 studies with a total of 2739 participants has shown that empathic accuracy is positively—though weakly—correlated with greater marital satisfaction as well (Sened et al., 2017). Interestingly, empathic accuracy for the partner's negative thoughts and feelings was more strongly correlated with marital satisfaction than empathic accuracy for the partner's positive thoughts and feelings, a finding which suggests that empathic accuracy may be particularly helpful when it is used to identify and address the partner's problems and concerns before small irritations and complaints escalate into larger ones (cf. Simpson, Ickes, & Oriña, 2001).

The deceptively simple conclusion that appears to follow from this research is that more understanding is good for relationships. Although this general claim has been endorsed by many practitioners and researchers, the overall association between empathic and relationship satisfaction ($r = 0.134$, $p < 0.05$) has proved to be considerably weaker than one might expect (Sened et al., 2017), and a number of authors have warned that greater understanding can have a destructive downside when partners come to understand things that have the potential to threaten and undermine their relationship (Bochner, 1981; Parks, 1981). Indeed, several studies have reinforced this concern by identifying conditions in which greater understanding is associated with more conflict and more relationship dissatisfaction, rather than less (e.g., Sillars, 1981, 1985; Sillars & Parry, 1982; Sillars, Pike, Jones, & Redmon, 1983; Sillars & Scott, 1983).

First, greater empathic accuracy might expose irreconcilable differences between the partners' perspectives, a condition in which further understanding will not lead to the convergence of these perspectives but will instead increase the levels of conflict or dissatisfaction.

Second, greater empathic accuracy might also reveal benign misconceptions that previously helped to maintain the relationship. When exposed as misconceptions, they have the potential to de-stabilize the relationship and to increase dissatisfaction.

Third, greater empathic accuracy might disclose blunt and harmful truths held by the target (e.g., concealment of an ongoing affair) that now present an immediate threat to the perceiver's self-image and to the relationship itself.

Viewed collectively, these three conditions can help us understand why the illusion of similarity (e.g., *projection*; Sillars, 1985; Sillars, Weisberg, Burggraf, & Zietlow, 1990; *assumed similarity*; Kenny, 1994) and the illusion of understanding (e.g., *perceived empathic effort*; Cohen, Schulz, Weiss, & Waldinger, 2012; *perceived understanding*; Hinnekens, Stas, Gistelinck, & Verhofstadt, 2020; Pollmann & Finkenauer, 2009) can at times be more important to relationship stability and satisfaction than actual understanding is.

To explain why greater empathic accuracy can often help, but sometimes hurt, close relationships, Ickes and Simpson (1997, 2001) have developed a nuanced model that takes into account both the partners' motives and the expected outcomes of empathic accuracy given the situation. The key variable in this model is whether the situation is one in which greater empathic accuracy is either likely or unlikely to reveal information that potentially threatens the partners' relationship.

The Empathic (In)Accuracy Model

Ickes and Simpson's (1997, 2001) empathic accuracy model proposes a phenomenon which they call *motivated inaccuracy*—a motivated suppression or restriction of empathic accuracy that occurs when the perceiver believes that the partner's actual thoughts and feelings are likely to threaten the relationship, the perceiver's self-esteem, or both. Simpson, Ickes, and Blackstone (1995) found compelling evidence for motivated inaccuracy in a study of dating couples. Interestingly, those couples who used motivated inaccuracy to avoid knowing each other's potentially threatening thoughts and feelings about attractive alternative dating partners were more likely than other couples to still be dating 5 months later, suggesting that "not going there" can be an effective strategy for sustaining the relationship during periods of relationship threat.

Findings from a subsequent study in sample of married couples were also consistent with the predictions of the Ickes and Simpson model: when the partner's thoughts and feelings were relationship-threatening, greater empathic accuracy was associated with a decline in the perceiver's feelings of closeness to the partner; the reverse was true when the partner's thoughts and feelings were nonthreatening (Simpson et al., 2003). On the other hand, recent study by Hinnekens et al. (2018) found only partial support for the model's predictions. A shift in participants' motivation to be accurate to a motivation to be inaccurate in response to perceived threat could not be detected in their sample of committed couples. For men, a higher level of empathic accuracy for the nonthreatening feelings of their female partner was predictive of an increased feeling of closeness, whereas for women it was predictive of a better mood. However, a harmful effect of empathic accuracy for threatening thoughts/feelings on personal or relationship well-being was not found.

Conclusion

Is greater empathic accuracy associated with greater marital satisfaction? As previous writers (Hinnekens et al., 2018; Ickes & Simpson, 2001; West, 2008) have noted, it is important to distinguish between short-term and long-term outcomes when measuring the impact of empathic accuracy on relational well-being. West (2008) argued that we need to know how partners compute their cost-benefit ratios when they decide how accurate to be in a given situation, and that we also need to know whether these decisions are driven by short-term or long-term motives. This argument implies that other considerations may account for the fact that studies examining the association between empathic accuracy and relationship satisfaction have yielded mixed results. One example of such a consideration is pragmatism; although the target's thoughts might be perceived as potentially threatening, accurately inferring them might still be judged to be the most pragmatic option available, and therefore required (e.g., in order to resolve the conflict, face recurring difficulties, or confront unwanted behavior). This is an important insight that deserves to be fully explored in future research.

Furthermore, not only pragmatism but other important moderators also need further research to comprehend the complex role of empathic accuracy in relationship well-being. For example, partners' mental state will definitely influence how accurate partners' perceive each other's inner world as psychopathology not only sets the boundaries for the cognitive resources and abilities one has to take other's perspective (see "***Empathic accuracy and psychopathology***"*)* but also determines one's emotional state that serves as the basis for interpersonal processes such as closeness or attunement—necessary when estimating the emotional state of the other. Research on the role of depressive symptoms in the context of marital conflict showed a significant interconnectedness between partners' affectivity during processes such as empathic accuracy and assumed similarity (Papp, Kouros, & Cummings, 2010). This study indicated that among couples with higher depressive symptoms, empathic accuracy for negative emotions such as anger or fear is lower and assumed similarity for these negative emotions is higher demonstrating that *negative* conflict expressions are of particular importance in the interplay between mental well-being and intimate relationships. Although perspective-taking processes may be hindered by psychopathology leading to a decrease in empathic accuracy (i.e., a specific momentary situational interaction-based measure), perceived understanding (i.e,.a global measure) may have a protective role as this subjective feeling may matter more for partners' relationship well-being as found by several studies that has taken different forms of understanding into account (e.g., Hinnekens et al., 2020; Pollmann & Finkenauer, 2009). Further research on the different aspects of understanding is therefore indispensable and certainly within the specific context of psychopathology within couples.

Family Relationships

One would expect that family members—with their longstanding joint history—will generally understand each other very accurately, because the shared knowledge that comes with long-term acquaintance can help us infer and interpret unspoken thoughts, ambiguous statements, or references to events that occurred at a different place and time (Stinson & Ickes, 1992). Research on empathic accuracy in families is scarce, but the available findings indicate that family members rate their own perceiving skills being as very accurate and that they feel confident about their understanding of the other family members. However, research conducted by Sillars, Koerner, and Fitzpatrick (2005) shows that these subjective ratings are in significant contrast to the perceivers' relatively low, objectively measured accuracy scores. The same mechanism underlying the negative correlation between relationship duration and empathic accuracy in marriage partners may also be at play here, as family members appear to select and interpret the available cues in an automatic way, either because they think "they already heard it before" or because they select only those cues that are consistent with their already-established partner-specific schemas (Sillars, 2011).

Also interesting is the role of age differences and the associated differences in cognitive development between family members, for example, between parents and their adolescent children. Parents often assume that their adolescents' thoughts are complex "process-thoughts" (i.e., thoughts concerning the course of the interaction or conveying underlying relational messages), whereas adolescents usually report thoughts that involve little or no subtext and are literally associated with the subject of the interaction (Sillars et al., 2005). This is another reason why, despite their long acquaintance based on a joint history of interactions and familiarity with each other, family members are, at best, only moderately accurate in "reading" each other's thoughts and feelings.

Empathic Accuracy in Clinical Settings

Empathic Accuracy in Distressed Relationships

Studies of distressed or poorly adjusted couples have found that the partners' perceptions of communication and their attributions about each other are biased and often incongruent with their intentions. These findings suggest that distressed partners are either unable to express their intended meanings or are biased in reporting their actual intentions (Gottman, Markman, & Notarius, 1977; Madden & Janoff-Bulman, 1981). In other words, the relationships of distressed couples tend to be characterized by important misperceptions and a lack of mutual understanding.

A similar finding has been documented in research using the empathic accuracy paradigm in the relationships of maritally abusive men. Schweinle, Ickes, and Bernstein (2002) found that these men are more likely than non-abusive men to disattend women's complaints and to inaccurately infer their thoughts and feelings. In a later study (Schweinle & Ickes, 2007), these findings were replicated and a cognitive bias which the authors called *the critical/rejecting overattribution bias* was identified in the perceptions of abusive men. These men interpreted their wives' thoughts about them as being more critical and rejecting than was actually the case, and the degree of this bias was associated with the degree of impairment in their empathic accuracy regarding their wives' thoughts and feelings. Interestingly, abusive men were significantly less accurate when they inferred their own wives' thoughts and feelings than when they inferred the thoughts and feelings of female strangers (Clements, Holtzworth-Munroe, Schweinle, & Ickes, 2007).

Empathic Accuracy and Psychopathology

Individuals diagnosed with an autism spectrum disorder often exhibit social limitations due to their inability to correctly assign mental states and feelings to themselves and to others, a limitation in the so-called "theory of mind" capacity. It has been further proposed that these individuals also fail to adequately recognize and infer the unspoken thoughts and feelings of others and, consequently, exhibit socially inadequate behavior (Baron-Cohen, Tager-Flusberg, & Cohen, 1994). Several diagnostic tests have been created to measure the level of development of individuals' theory of mind capacities, such as the Reading the Mind in the Eyes test. However, because the complexities of social interactions are not represented in these static tests, the empathic accuracy paradigm provides a more natural and ecologically valid test.

When normally developing research participants are compared with participants diagnosed with a pervasive developmental disorder (PDD; Roeyers, Buysse, Ponnet, & Pichal, 2001) or with a mild degree of autism (Demurie, De Corel, & Roeyers, 2011), the PDD and mildly autistic participants had significantly lower empathic accuracy scores, as measured in studies using the SS-paradigm. This difference was particularly evident in unstructured conversations, suggesting that a predictable structure can help PDD and mildly autistic individuals to partially compensate for their disability (Ponnet et al., 2008).

Empathic Accuracy in Clinicians

Accurately assessing thoughts and feelings of clients is a key factor in clinical practice and is frequently cited as one of the most important predictors of successful client-centered psychotherapy (Greenberg, Watson, Elliot, & Bohart, 2001). Carl Rogers suggested that accurate empathy—in addition to the therapist's authenticity and unbiased care for the client—is a necessary attribute of a good therapist (Rogers, 1957). Other schools of psychotherapy have similarly emphasized the importance of perspective-taking skills, which they define as the ability of the therapist to perceive self and others as social beings with subjective states of mind and internal mental processes (e.g., mentalization-based treatment; Allen & Fonagy, 2006). Consistent with these views, Greenberg et al. (2001) found in their meta-analysis that the degree of the therapist's empathy—defined as: *[Empathy is] the therapist's sensitive ability and willingness to understand the client's thoughts, feelings, and struggles from the client's point of view* (Rogers, 1980, p. 85)—was more important for the outcome of the psychotherapy than were specific interventions (i.e., interventions within the specific theoretical orientation of the therapist). It is important to note that not only the client's perception of being understood was predictive, but also the objective accuracy score of the therapist (i.e., observer ratings of the therapist's empathy). In addition, the link between therapist empathy and therapy outcomes was the strongest in young and less experienced psychotherapists, presumably because they vary more in empathic skill. The research of Kwon and Jo (2012) further confirms that as clinicians become more experienced, their empathic accuracy increases, reducing its variability across therapists and enabling other, more variable skills (e.g., effective problem-solving) to emerge as significant predictors of positive therapy outcomes.

Trainability of Empathic Accuracy in Clinicians

As Ickes and colleagues suggest, the research on empathic accuracy offers a methodology that can be used, not just to objectively measure, but also to train a perceiver's cognitive empathy (Ickes, 2003; Schmid Mast & Ickes, 2007). Marangoni et al. (1995) showed that, when undergraduate perceivers were asked to infer the thoughts and feelings of female targets who discussed their relationship problems with a male, client-centered therapist, their empathic accuracy improved over the course of each client's videotape. Of even greater interest, the perceivers who received immediate feedback about the accuracy of their inferences during the middle portion of each tape improved their accuracy even more, resulting in an estimated 10% greater increase than was observed for perceivers in the no-feedback control condition.

These results suggest that actual psychotherapists also need time to get to know the client well enough to achieve higher levels of EA, regardless of the initial differences in expertise of the therapist. In addition, these results demonstrate a way to further train and enhance the empathic skills of clinicians. Important to note is that although the training effect seemed to generalize over the three different targets, it was somewhat less effective for a client who expressed many ambivalent thoughts and feelings.

To test for the effect of feedback training in actual student therapists, Barone et al. (2005) conducted an intervention-training study in which final-year clinical psychology students observed therapeutic sessions and inferred the thoughts and feelings of clients. Compared to a control group, the students who received the feedback intervention show significantly greater accuracy for the clients' feelings—but not for their thoughts. The authors speculated that the training effect was found only for feelings because the thought-content in clinical subjects is often unusual and reflective of their psychopathology, and therefore more difficult to infer. In addition, it is also possible that one can make faster, more accurate estimates about feelings because they are more limited in number in contrast to the unlimited range of potential thoughts.

Empathic Accuracy in Couple Therapy

Another important question is whether partner accuracy matters in couple therapy, because a common complaint of partners seeking marital help is a lack of mutual understanding in their relationship (Laing et al., 1966). Clinicians often advise the partners to express their thoughts and feelings more openly to each other in order to increase their levels of mutual understanding. However logical this assumption might seem, research on the association between open, direct communication and empathic accuracy has yielded conflicting and counterintuitive results. For example, in one study there was no association between the perceiver's empathic accuracy and the extent to which the other partner expressed his or her thoughts and feelings openly and clearly in behavior (i.e., in a measure of "readability" like the type described earlier; Thomas et al., 1997). On the other hand, in research with dating partners and interaction partners who did not know each other, a positive relationship between "readability" and empathic accuracy was found (Simpson et al., 1995; Stinson & Ickes, 1992; Thomas & Fletcher, 2003). This phenomenon can be explained, as we have noted earlier, by the fact that long-term partners may no longer rely on immediate behavioral cues to make their empathic inferences, instead relying on the partner-specific schemas they have developed. Even more surprising, the frequency of which partners discuss a particular topic contributes little to nothing to the mutual understanding the partners achieve with regard to these conflict issues (Sillars et al., 1994). Quite ironically, the less constructive forms of communication—in particular, expressed negativity (e.g., verbal competitiveness, negative intonation)—predict a better mutual understanding because it provides a clear and unambiguous signal of dissatisfaction regarding the topic.

As discussed above, empathic accuracy not only increases over time but can be further increased through feedback training. Accuracy levels in friendships or young partnerships will initially increase as a result of the partners' increased exposure to each other, and there is additional room for improvement if the other person gives a lot of immediate feedback about the correctness of one's empathic inferences. This implies that empathic accuracy should also increase when one explicitly "tests" the accuracy of one's inferences (e.g., "Do I understand you correctly by saying that you …"; "Is it true that you are worried about this?"). The feedback from the target partner teaches something about the connection between the outside (the verbal and non-verbal signals) and the inside (thoughts and feelings) of the other. Having the partners actively solicit such immediate feedback from each other is therefore a promising therapeutic tool that should be discussed with, and practiced by, the clients.

In addition, clinicians can alert couples and family to those automatic thoughts and cognitive biases that can limit or undermine their ability to correctly "read" each other (for example, "All of her thoughts about me are critical and rejecting," "I already know what he's thinking and therefore don't have to pay any attention to what he is saying," and "His thoughts aren't interesting enough for me to try to figure out"). These automatic thoughts and cognitive biases often lead to premature judgments that are not based on the objective information that the other person actually transmits, but instead on the automatic and stable partner-specific schemas that the perceiver has formed. To be sure, such rapid and automatic processing can help us in everyday life to respond quickly and effectively without having to exert much effort. However, these cognitive shortcuts can also lead to misunderstanding and, in the long term, even to perpetual feelings of incomprehension and of "not being heard." Effective feedback training that actively involves both partners who are committed to understanding each other's actual thoughts and feelings may be essential to correcting the faulty, and ultimately lazy, mind-reading habits that characterize many, if not most, distressed relationships.

References

Allen, J. G., & Fonagy, P. (2006). *The handbook of mentalization-based treatment*. Chichester: John Wiley & Sons.

Baron-Cohen, S., Tager-Flusberg, H., & Cohen, D. J. (1994). *Understanding other minds: Perspectives from autism*. New York, NY: Oxford University Press.

Barone, D. F., Hutchings, P. S., Kimmel, H. J., Traub, H. L., Cooper, J. T., & Marshall, C. M. (2005). Increasing empathic accuracy through practice and feedback in a clinical interviewing course. *Journal of Social and Clinical Psychology, 24*, 156–171. https://doi.org/10.1521/jscp.24.2.156.62275

Bochner, A. P. (1981). On the efficacy of openness in close relationships. In M. Burgoon (Ed.), *Communication yearbook* (pp. 109–124). New Brunswick, NJ: Transaction Books.

Bodenmann, G. (2005). Dyadic coping and its significance for marital functioning. In T. A. Revenson, K. Kayser, & G. Bodenmann (Eds.), *Couples coping with stress: Emerging perspectives on dyadic coping* (pp. 33–49). Washington, DC: American Psychological Association.

Christensen, L., & Wallace, L. (1976). Perceptual accuracy as a variable in marital adjustment. *Journal of Sex & Marital Therapy, 2*, 130–136. https://doi.org/10.1080/00926237608402971

Clements, K., Holtzworth-Munroe, A., Schweinle, W., & Ickes, W. (2007). Empathic accuracy of intimate partners in violent versus nonviolent relationships. *Personal Relationships, 14*, 369–388. https://doi.org/10.1111/j.1475-6811.2007.00161.x

Cohen, S., Schulz, M. S., Weiss, E., & Waldinger, R. J. (2012). Eye of the beholder: The individual and dyadic contributions of empathic accuracy and perceived empathic effort to relationship satisfaction. *Journal of Family Psychology, 26*, 236–245. https://doi.org/10.1037/a0027488

Corsini, R. J. (1956). Understanding and similarity in marriage. *The Journal of Abnormal and Social Psychology, 52*, 327–332. https://doi.org/10.1037/h0043556

Costanzo, M., & Archer, D. (1989). Interpreting the expressive behavior of others: The interpersonal perception task. *Journal of Nonverbal Behavior, 13*, 225–245. https://doi.org/10.1007/BF00990295

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*, 113–126. https://doi.org/10.1037/0022-3514.44.1.113

Demurie, E., De Corel, M., & Roeyers, H. (2011). Empathic accuracy in adolescents with autism spectrum disorders and adolescents with attention-deficit/hyperactivity disorder. *Research in Autism Spectrum Disorders, 5*, 126–134. https://doi.org/10.1016/j.rasd.2010.03.002

Devoldre, I., Davis, M., Verhofstadt, L. L., & Buysse, A. (2010). Empathy and social support provision in couples: Social support and the need to study the underlying processes. *The Journal of Psychology, 144*, 259–284. https://doi.org/10.1080/00223981003648294

Dugosh, J. W. (2001). *Effects of relationship threat and ambiguity on empathic accuracy in dating couples*. Unpublished doctoral dissertation, University of Texas at Arlington.

Dymond, R. (1954). Interpersonal perception and marital happiness. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 8*, 164–171. https://doi.org/10.1037/h0083611

Ferguson, L. R., & Allen, D. R. (1978). Congruence of parental perception, marital satisfaction, and child adjustment. *Journal of Consulting and Clinical Psychology, 46*, 345–346. https://doi.org/10.1037/0022-006X.46.2.345

Gesn, P. R., & Ickes, W. (1999). The development of meaning contexts for empathic accuracy: Channel and sequence effects. *Journal of Personality and Social Psychology, 77*, 746–761. https://doi.org/10.1037/0022-3514.77.4.746

Gottman, J., Markman, H., & Notarius, C. (1977). The topography of marital conflict: A sequential analysis of verbal and nonverbal behavior. *Journal of Marriage and the Family, 39*, 461–477. https://doi.org/10.2307/350902

Greenberg, L. S., Watson, J. C., Elliot, R., & Bohart, A. C. (2001). Empathy. *Psychotherapy: Theory, Research, Practice, Training, 38*, 380–384. https://doi.org/10.1037/0033-3204.38.4.380

Gurman, A. S. (2008). A framework for the comparative study of couple therapy. In A. S. Gurman (Ed.), *Clinical handbook of couple therapy* (pp. 1–26). New York, NY: Guilford Press.

Guthrie, D. M., & Noller, P. (1988). Spouses' perceptions of one another in emotional situations. In P. Noller & M. A. Fitzpatrick (Eds.), *Perspectives on marital interaction* (pp. 53–181). Clevedon: Multilingual Matters.

Hall, J. A. (1984). *Nonverbal sex differences: Communication accuracy and expressive style*. Baltimore, MD: The Johns Hopkins University Press.

Hall, J. A., Blanch, D. C., Horgan, T. G., Murphy, N. A., Rosip, J. C., & Schmid Mast, M. (2009). Motivation and interpersonal sensitivity: Does it matter how hard you try? *Motivation and Emotion, 33*, 291–302. https://doi.org/10.1007/s11031-009-9128-2

Hall, J. A., & Schmid Mast, M. (2007). Sources of accuracy in the empathic accuracy paradigm. *Emotion, 7*, 438–446. https://doi.org/10.1037/1528-3542.7.2.43

Hinnekens, C. (2017). *Empathic (in)accuracy during couples' conflict interactions.* Unpublished doctoral dissertation, Ghent University, Faculty of Psychology and Educational Sciences, Ghent, Belgium.

Hinnekens, C., Ickes, W., De Schryver, M., & Verhofstadt, L. (2016). Demand behavior and empathic accuracy in observed conflict interactions in couples. *Journal of Social Psychology, 156*, 437–443. https://doi.org/10.1080/00224545.2015.1115386

Hinnekens, C., Loeys, T., De Schryver, M., & Verhofstadt, L. L. (2018). The manageability of empathic (in)accuracy during couples' conflict: Relationship-protection or self-protection? *Motivation and Emotion, 42*, 403–418. https://doi.org/10.1007/s11031-018-9689-z

Hinnekens, C., Stas, L., Gistelinck, F., & Verhofstadt, L. L. (2020). "I think you understand me." Studying the associations between actual, assumed, and perceived understanding within couples. *European Journal of Social Psychology, 50*, 46–60. https://doi.org/10.1002/ejsp.2614

Hodges, S. D., Laurent, S. M., & Lewis, K. L. (2011). Specially motivated, feminine, or just female: Do women have an empathic accuracy advantage? In J. Smith, W. Ickes, J. Hall, & S. Hodges (Eds.), *Managing interpersonal sensitivity: Knowing when & when not to understand others* (pp. 59–74). Hauppauge, NY: Nova Science.

Hodges, S. D., Lewis, K. L., & Ickes, W. (2015). The matter of other minds: Empathic accuracy and the factors that influence it. In P. Shaver & M. Mikulincer (Eds.), *APA handbook of personality and social psychology: Vol 3. Interpersonal relations and group processes* (pp. 319–348). Washington, DC: American Psychological Association.

Ickes, W. (1982). A basic paradigm for the study of personality, roles, and social behavior. In W. Ickes & E. S. Knowles (Eds.), *Personality, roles, and social behavior* (pp. 305–341). New York, NY: Springer.

Ickes, W. (1993). Empathic accuracy. *Journal of Personality, 61*, 587–610. https://doi.org/10.1111/j.1467-6494.1993.tb00783.x

Ickes, W. (1997). *Empathic accuracy*. New York, NY: Guilford Press.

Ickes, W. (2003). *Everyday mind reading: Understanding what other people think and feel*. Amherst, NY: Prometheus Books.

Ickes, W. (2011). Everyday mind reading is driven by motives and goals. *Psychological Inquiry, 22*, 200–206. https://doi.org/10.1080/1047840X.2011.561133

Ickes, W., Buysse, A., Pham, H., Rivers, K., Erickson, J. R., Hancock, M., … Gesn, P. R. (2000). On the difficulty of distinguishing "good" and "poor" perceivers: A social relations analysis of empathic accuracy data. *Personal Relationships, 7*, 219–234. https://doi.org/10.1111/j.1475-6811.2000.tb00013.x

Ickes, W., Gesn, P. R., & Graham, T. (2000). Gender differences in empathic accuracy: Differential ability or differential motivation? *Personal Relationships, 7*, 95–109. https://doi.org/10.1111/j.1475-6811.2000.tb00006.x

Ickes, W., & Hodges, S. D. (2013). Empathic accuracy in close relationships. In J. A. Simpson & L. Campbell (Eds.), *The Oxford handbook of close relationships* (pp. 348–373). Oxford: Oxford University Press.

Ickes, W., & Simpson, J. A. (1997). Managing empathic accuracy in close relationships. In W. Ickes (Ed.), *Empathic accuracy* (pp. 218–250). New York, NY: Guilford Press.

Ickes, W., & Simpson, J. A. (2001). Motivational aspects of empathic accuracy. In G. J. O. Fletcher & M. S. Clark (Eds.), *Interpersonal processes: Blackwell handbook in social psychology* (pp. 229–249). Oxford: Blackwell.

Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). Naturalistic social cognition: Empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology, 59*, 730–742. https://doi.org/10.1037/0022-3514.59.4.730

Ickes, W., & Tooke, W. (1988). The observational method: Studying the interaction of minds and bodies. In S. Duck, D. F. Hay, S. E. Hobfoll, W. Ickes, & B. M. Montgomery (Eds.), *Handbook of personal relationships: Theory, research and interventions* (pp. 79–97). Oxford: John Wiley & Sons.

Kagan, N. (1977). *Interpersonal process recall*. East Lansing, MI: Michigan State University Press.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. London: Guilford Press.

Kilpatrick, S. D., Bissonnette, V. L., & Rusbult, C. E. (2002). Empathic accuracy and accommodative behavior among newly married couples. *Personal Relationships, 9*, 369–393. https://doi.org/10.1111/1475-6811.09402

Klein, K. J. K., & Hodges, S. D. (2001). Gender differences, motivation and empathic accuracy: When it pays to understand. *Personality and Social Psychology Bulletin, 27*, 720–730. https://doi.org/10.1177/0146167201276007

Kwon, K. I., & Jo, S. Y. (2012). The relationship among counselor experience level, empathic accuracy, and counseling outcome in the early phase of counseling. *Asia Pacific Education Review, 13*, 771–777. https://doi.org/10.1007/s12564-012-9235-8

Laing, R. D., Phillipson, H., & Lee, A. R. (1966). *Interpersonal perception: A theory and a method of research*. Oxford: Springer.

Leary, M. R., Kelly, K.M., & Schreindorfer, L. S. (2001). *Individual differences in the need to belong*. Unpublished manuscript, Wake Forest University, Winston-Salem, NC.

Lewis, K. L. & Hodges, S. D. (2009). *Empathic accuracy and nonverbal decoding: Related or distinct constructs?* Unpublished data, University of Oregon, Eugene.

Luckey, E. B. (1960). Marital satisfaction and congruent self-spouse concepts. *Social Forces, 39*, 153–157. https://doi.org/10.2307/2574154

Madden, M. E., & Janoff-Bulman, R. (1981). Blame, control, and marital satisfaction: Wives' attributions for conflict in marriage. *Journal of Marriage and the Family, 43*, 663–674. https://doi.org/10.2307/351767

Marangoni, C., Garcia, S., Ickes, W., & Teng, G. (1995). Empathic accuracy in a clinically relevant setting. *Journal of Personality and Social Psychology, 68*, 854–869. https://doi.org/10.1037/0022-3514.68.5.854

McLeod, J. M., & Chaffee, S. H. (1973). Interpersonal approaches to communication research. *American Behavioral Scientist, 16*, 469–499. https://doi.org/10.1177/000276427301600402

Mehrabian, A., & Epstein, N. (1972). A measure of emotional empathy. *Journal of Personality, 40*, 525–543. https://doi.org/10.1111/j.1467-6494.1972.tb00078.x

Nowicki, S., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The diagnostic analysis of nonverbal accuracy scale. *Journal of Nonverbal Behavior, 18*, 9–35. https://doi.org/10.1007/BF02169077

Papp, L. M., Kouros, C. D., & Cummings, E. M. (2010). Emotions in marital conflict interactions: Empathic accuracy, assumed similarity, and the moderating context of depressive symptoms. *Journal of Social and Personal Relationships, 27*, 367–387. https://doi.org/10.1177/0265407509348810

Parks, M. R. (1981). Ideology in interpersonal communication. Off the couch and into the world. In M. Burgoon (Ed.), *Communication yearbook* (Vol. 5, pp. 79–108). New Brunswick, NJ: Transaction Books.

Pickett, C. L., Gardner, W. L., & Knowles, M. (2004). Getting a cue: The need to belong and enhanced sensitivity to social cues. *Personality and Social Psychology Bulletin, 30*, 1095–1107. https://doi.org/10.1177/0146167203262085

Pollmann, M. M., & Finkenauer, C. (2009). Investigating the role of two types of understanding in relationship well-being: Understanding is more important than knowledge. *Personality and Social Psychology Bulletin, 35*, 1512–1527. https://doi.org/10.1177/0146167209342754

Ponnet, K., Buysse, A., Roeyers, H., & De Clercq, A. (2008). Mind-reading in young adults with ASD: Does structure matter? *Journal of Autism and Developmental Disorders, 38*, 905–918. https://doi.org/10.1007/s10803-007-0462-5

Ponnet, K., Roeyers, H., Buysse, A., De Clercq, A., & Van Der Heyden, E. (2004). Advanced mind-reading in adults with Asperger syndrome. *Autism, 8*, 249–266. https://doi.org/10.1177/1362361304045214

Roeyers, H., Buysse, A., Ponnet, K., & Pichal, B. (2001). Advancing advanced mind-reading tests: Empathic accuracy in adults with a pervasive developmental disorder. *The Journal* *of Child Psychology and Psychiatry and Allied Disciplines, 42*, 271–278. https://doi.org/10.1111/1469-7610.00718

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology, 21*, 95–103. https://doi.org/10.1037/h0045357

Rogers, C. R. (1980). *A way of being*. Boston, MA: Houghton Mifflin.

Rusbult, C. E., Verette, J., Whitney, G. A., Slovik, L. F., & Lipkus, I. (1991). Accommodation processes in close relationships: Theory and preliminary empirical evidence. *Journal of Personality and Social Psychology, 60*, 53–78. https://doi.org/10.1037/0022-3514.60.1.53

Schmid Mast, M., & Ickes, W. (2007). Empathic accuracy: Measurement and potential clinical applications. In T. F. D. Farrow & P. W. R. Woodruff (Eds.), *Empathy in mental illness* (pp. 408–427). Cambridge, MA: Cambridge University Press.

Schweinle, W. E., & Ickes, W. (2007). The role of men's critical/rejecting overattribution bias, affect, and attentional disengagement in marital aggression. *Journal of Social and Clinical Psychology, 26*, 173–198. https://doi.org/10.1521/jscp.2007.26.2.173

Schweinle, W. E., Ickes, W., & Bernstein, I. H. (2002). Emphatic inaccuracy in husband to wife aggression: The over attribution bias. *Personal Relationships, 9*, 141–158. https://doi.org/10.1111/1475-6811.00009

Sened, H., Lavidor, M., Lazarus, G., Bar-Kalifa, E., Rafaeli, E., & Ickes, W. (2017). Empathic accuracy and relationship satisfaction: A meta-analytic review. *Journal of Family Psychology, 31*, 742–752. https://doi.org/10.1037/fam0000320

Sillars, A., & Scott, M. (1983). Interpersonal perception between intimates: An integrative review. *Human Communication Research, 10*, 153–176. https://doi.org/10.1111/j.1468-2958.1983.tb00009.x

Sillars, A. L. (1981). Attributions and interpersonal conflict resolution. In J. H. Harvey, W. J. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research* (Vol. 3, pp. 279–305). Hillsdale, NJ: Erlbaum.

Sillars, A. L. (1985). Interpersonal perception in relationships. In W. J. Ickes (Ed.), *Compatible and incompatible relationships* (pp. 277–305). New York, NY: Springer.

Sillars, A. L. (2011). Motivated misunderstanding in family conflict discussions. In J. L. Smith, W. Ickes, J. Hall, & S. Hodges (Eds.), *Managing interpersonal sensitivity: Knowing when – And when not – To understand others* (pp. 193–213). Hauppauge, NY: Nova Science.

Sillars, A. L., Folwell, A. L., Hill, K. C., Maki, B. K., Hurst, A. P., & Casano, R. A. (1994). Marital communication and the persistence of misunderstanding. *Journal of Social and Personal Relationships, 11*, 611–617. https://doi.org/10.1177/0265407594114008

Sillars, A. L., Koerner, A., & Fitzpatrick, M. A. (2005). Communication and understanding in parent–adolescent relationships. *Human Communication Research, 31*, 102–128. https://doi.org/10.1111/j.1468-2958.2005.tb00866.x

Sillars, A. L., & Parry, D. (1982). Stress, cognition and communication in interpersonal conflicts. *Communication Research, 9*, 201–226. https://doi.org/10.1177/009365082009002002

Sillars, A. L., Pike, G. R., Jones, T. S., & Redmon, K. (1983). Communication and conflict in marriage. In R. Bostrom (Ed.), *Communication yearbook* (Vol. 7, pp. 414–429). Beverly Hills, CA: Sage.

Sillars, A. L., Weisberg, J., Burggraf, C. S., & Zietlow, P. H. (1990). Communication and understanding revisited: Married couples' understanding and recall of conversations. *Communication Research, 17*, 500–522. https://doi.org/10.1177/009365090017004006

Simpson, J. A., Ickes, W., & Blackstone, T. (1995). When the head protects the heart: Empathic accuracy in dating relationships. *Journal of Personality and Social Psychology, 69*, 629–641. https://doi.org/10.1037/0022-3514.69.4.629

Simpson, J. A., Ickes, W., & Oriña, M. (2001). Empathic accuracy and preemptive relationship maintenance. In J. H. Harvey & A. Wenzel (Eds.), *Close romantic relationships: Maintenance and enhancement* (pp. 27–46). Mahwah, NJ: Lawrence Erlbaum Associates.

Simpson, J. A., Kim, J. S., Fillo, J., Ickes, W., Rholes, S., Oriña, M. M., & Winterheld, H. A. (2011). Attachment and the management of empathic accuracy in relationship threatening situations. *Personality and Social Psychology Bulletin, 37*, 242–254. https://doi.org/10.1177/0146167210394368

Simpson, J. A., Oriña, M. M., & Ickes, W. (2003). When accuracy hurts, and when it helps: A test of the empathic accuracy model in marital interactions. *Journal of Personality and Social Psychology, 85*, 881–893. https://doi.org/10.1037/0022-3514.85.5.881

Smith, J. L., Ickes, W., Hall, J. A., & Hodges, S. (2011). *Managing interpersonal sensitivity: Knowing when & when not to understand others*. Hauppauge, NY: Nova Science.

Stinson, L., & Ickes, W. (1992). Empathic accuracy in the interactions of male friends versus male strangers. *Journal of Personality and Social Psychology, 62*, 787–797. https://doi.org/10.1037/0022-3514.62.5.787

Stuckert, R. P. (1963). Role perception and marital satisfaction. A configurational approach. *Marriage and Family Living, 25*, 415–419. https://doi.org/10.2307/349038

Thomas, G., & Fletcher, G. J. (2003). Mind-reading accuracy in intimate relationships: Assessing the roles of the relationship, the target, and the judge. *Journal of Personality and Social Psychology, 85*, 1079–1094. https://doi.org/10.1037/0022-3514.85.6.1079

Thomas, G., Fletcher, G. J. O., & Lange, C. (1997). On-line empathic accuracy in marital interaction. *Journal of Personality and Social Psychology, 72*, 839–850. https://doi.org/10.1037/0022-3514.72.4.839

Verhofstadt, L. L., Buysse, A., Ickes, W., Davis, M., & Devoldre, I. (2008). Support provision in marriage: The role of emotional similarity and empathic accuracy. *Emotion, 8*, 792–802. https://doi.org/10.1037/a0013976

Verhofstadt, L. L., Devoldre, I., Buysse, A., Stevens, M., Hinnekens, C., Ickes, W., & Davis, M. (2016). The role of cognitive and affective empathy in spouses' support interactions: An observational study. *PLoS One, 11*, e0149944. https://doi.org/10.1371/journal.pone.0149944

Wegner, D. M., Giuliano, T., & Hertel, P. T. (1985). Cognitive interdependence in close relationships. In W. Ickes (Ed.), *Compatible and incompatible relationships* (pp. 253–276). New York, NY: Springer.

West, T. W. (2008). *Four principles in the study of bias and accuracy in close relationships*. Ann Arbor, MI: ProQuest.

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological Science, 19*, 399–404. https://doi.org/10.1111/j.1467-9280.2008.02099.x

Empathic Accuracy: Lessons from the Perception of Contextualized Real-Life Emotional Expressions

Doron Atias and Hillel Aviezer

Introduction

A child is insulted by his peers during play and runs to his father. In a split second, Dad feels his child's pain and rushes to him, offering a comforting hug. Empathy, our ability to feel other's emotions is a fundamental facet of human behavior (Zaki & Ochsner, 2011), key for building and maintaining healthy social relationships. But how do we gain access to the internal feelings of others? One intuitive mechanism for solving this puzzle involves the accurate readout of affective cues from the overt emotional expressions of others (Ekman, 1993). Indeed, accurate recognition of emotional faces has been shown to predict empathic, prosocial behavior (Marsh, Kozak, & Ambady, 2007). Conversely, poor emotion recognition skills have been proposed as a causal mechanism leading to impaired empathy such as in the case of individuals with autism (Clark, Winkielman, & McIntosh, 2008). Emotion recognition may thus be considered a first, critical step to empathy.

The ability to draw correct inferences on the internal state of social others was measured and quantified extensively over the past decades using the paradigm of *empathic accuracy* (Ickes, Stinson, Bissonnette, & Garcia, 1990; Zaki, Bolger, & Ochsner, 2008). A typical administration of this paradigm contains three principal elements. First, social targets are requested to discuss emotional events from their lives, while being videotaped by the experimenter. Then, targets watch their own videos and supply an affective ground-truth report of their feelings throughout the video. Finally, naïve perceivers are asked to watch the videos and assess the feelings expressed by the targets in the videos. The correspondence between perceiver ratings and targets ratings constitutes the measurement of empathic accuracy.

This interpersonal approach for measuring empathic accuracy highlights not only the perceiver's ability to experience and assess targets' emotions, but also the

D. Atias (✉) · H. Aviezer
Department of Psychology, The Hebrew University of Jerusalem, Jerusalem, Israel
e-mail: doron.atias@mail.huji.ac.il; haviezer@huji.ac.il

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_9

targets tendency to express valid and diagnostic affective cues for the perceivers to read. Accordingly, it has been argued that perceivers' empathic accuracy is mediated by targets' expressivity. For example, perceivers that are high in trait affective empathy (as measured by perceivers' self-reports) show increased performance in empathic accuracy but only when targets expressivity is high (Zaki et al., 2008). Therefore, the diagnostic quality of emotional expressions conveyed by social targets plays a critical role in empathic accuracy.

The above assumptions raise a key question: are emotional expressions truly informative for affect judgment? As next reviewed, this is a subject of prolonged debate among researchers in the field of emotion recognition. Here, we cast doubt on the diagnostic nature of emotional expressions in light of recent findings that challenge traditional models of emotion. These findings present critical theoretical and methodological implications to the fields of both emotion recognition and empathic accuracy.

The Diagnostic Nature of Emotional Expressions: Traditional Models

Modern studies of emotional expressions were strongly inspired by the work of Charles Darwin's *Expression of emotions in man and animals (*1872*)*. Before Darwin, human facial expressions of emotion were considered God-given and divine, demonstrating the unique ability of humans to convey their affective state to others (reviewed in Fridlund, 1994). To rebut this theological standpoint, Darwin sought for communalities among expressive movements across species and documented facial expressions and gestures in humans, primates and other animals, arguing for cross-cultural and cross-species similarities (Fridlund, 1994; Russell, Bachorowski, & Fernández-Dols, 2003). Inspired by Darwin, scientists like Schlosberg (1941), Tomkins (1962) and their successors began exploring the nature of emotional expressions. Their diverse interpretations of Darwin's original work generated a rich theoretical and empirical landscape aimed to unfold the mechanism and function of emotional expressions. In this section, we provide a brief and selective overview of the leading classic theories of emotion. Specifically, we focus on each theory's view on the architecture of emotional expressions.

Basic Emotion Theory

The basic emotion theory derived mostly from the writings of Tomkins (1962, 1963), who postulated an innate system of discrete ("basic") emotions that are universal, biologically based, and instinctively expressed and identified from facial movements. This idea was advanced by Ekman's model of basic emotions that posits a fixed number of primary emotions, each supported by a distinct process called an “affect program” (Ekman, Friesen, & Ellsworth, 1972). These programs, when triggered, are responsible for eliciting the various components of an emotion, including its distinct physiological pattern, feelings, behavioral tendencies, and specific prototypical patterns of evoked facial and vocal expressions. For example, in response to an emotional elicitor that activates the “fear program,” a specific stereotypical neuromuscular activation of a fearful face configuration would arise. The fearful expresser then signals diagnostic and veridical information about her internal state in a biologically hardwired manner. In turn, the receiver is biologically prepared to “read” this information correctly, so that the message encoded by the expresser is accurately decoded (Ekman, 1993; Smith, Cottrell, Gosselin, & Schyns, 2005). While Ekman’s original work focused on the face, following research extended this approach to the voice (Sauter, Eisner, Ekman, & Scott, 2010, 2015) and to the body as well (de Gelder, 2006, 2009; Tracy & Matsumoto, 2008; Tracy & Robins, 2004).

The list of primary emotions varies across studies and recent accounts suggest a diverse range of universally recognized expressions (Keltner, Sauter, Tracy, & Cowen, 2019). However, a defined set of six basic emotions: happiness, surprise, fear, disgust, anger, and sadness are by far the most studied. This is best reflected in the many sets of emotional stimuli portraying posed and stereotypical expressions of emotions, tools that influenced literally thousands of studies in the field (Belin, Fillion-Bilodeau, & Gosselin, 2008; de Gelder & Van den Stock, 2011; Ekman, 1976; Hawk, Van der Schalk, & Fischer, 2008). Thus, a fundamental argument of basic emotion theory is that emotion categories are expressed stereotypically and distinctly in the face, voice, and body, and recognized automatically and universally by perceivers (Ekman & Rosenberg, 1997; Sauter et al., 2010; Tracy & Robins, 2008).

Appraisal Theories of Emotion

Basic emotion views strongly emphasize the stereotypical output of the emotion process—the emotional expression. By contrast, appraisal views of emotion have focused on the underlying cognitive appraisal processes that form the building blocks for driving emotions and expressions. According to appraisal theory (e.g., Roseman, 1984; Scherer, 1984; Smith & Ellsworth, 1985), emotions are episodes of temporary synchronizations of five interrelated components: cognitive appraisal, physiological response, experienced feeling, motivational action tendencies, and motor expression (Banse & Scherer, 1996). The component process model of emotion specifies that affective situations are appraised for (1) novelty, (2) intrinsic pleasantness, (3) goal attainment, (4) power/control, and (5) compatibility with standards (Scherer, 2009). With regard to emotional expressions, it has been suggested that every sequential stage in the appraisal process (termed, appraisal checks) evokes specific action units in the face or acoustic characteristics in the voice (Scherer, Mortillaro, & Mehu, 2017). The cumulative effect of these appraisal checks forms the final expression. Emotions are communicated efficiently because perceivers can recognize the affective state of the target by inferring the appraisals from the expressions.

Appraisal theories differ from basic emotions with regard to predictions of emotional expressions. First, because people exhibit different combinations of appraisal checks, the theory allows for much larger variability in the everyday facial and vocal reactions that people display. In particular, appraisal theorists do not limit themselves to a predefined list of specific stereotypical basic emotions (Scherer & Moors, 2019). Furthermore, the same muscular activity may occur in different emotional situations because both share an underlying appraisal check. For example, anger and concentration may both involve brow lowering because both share the appraisal check of "goal obstruction." Thus, appraisal theories argue that facial and vocal expressions are diagnostic of affective experience, because they are diagnostic for the specific appraisal checks that the individual used.

Dimensional Theories of Emotion

If basic emotion models focus on stereotypical facial configurations, and appraisal models focus on face activity reflecting appraisal checks, dimensional models further reduce emotion and expression to their most fundamental elements. In the most popular of these models, all affective states arise from two fundamental neurophysiological systems, one related to valence (a pleasure–displeasure continuum) and the other to arousal, or alertness (Posner, Russell, & Peterson, 2005; Russell, 1980). The origins of this approach can be traced back to Woodworth and Schlosberg (1938) and Schlosberg's (1952) work on photographed faces. In a series of factor-analytic studies, they demonstrated that facial expressions are not discretely independent from one another, but rather organized in a circular, interrelated dimensional arrangement. This approach was continued with Russell's (1980) circumplexed model of emotion, according to which affective experiences can be explained using a set of two primary dimensions: valence and arousal. These dimensions are proposed to be the building blocks of emotion representation, and so, each specific emotion can be defined as some combination of its valence and arousal constituents (Russell, 1980).

Based on this view, Russell and Bullock (1986) suggested a two-stage model of emotion perception. In the first stage, valence and arousal are rapidly and effortlessly read out directly from the sensory cues (be they a facial expression or a vocalization). While the sensory cue is sufficient to inform the perceiver if the affect is positive or negative (or aroused vs. deactivated), the combination of valence and arousal is "fuzzy" and overlapping with regard to specific discrete emotions (Russell & Bullock, 1986). Only later at a second stage, dimensional values are interpreted and transformed into more specific emotional labels using contextual information (Russell, 1997). Contemporary theories stemming from this approach have highlighted the importance of construction processes in the experience (Lindquist & Barrett, 2008) and perception (Lindquist & Gendron, 2013) of emotion. In particular, language has been shown to shape emotion perception of specific categories (Gendron, Lindquist, Barsalou, & Barrett, 2012; Lindquist & Gendron, 2013). Thus, according to dimensional views of emotion, emotional expressions are diagnostic for affective valence and arousal experienced by the target.

Shared Perspectives in Classic Emotion Theories

Our brief and selective overview of classic emotion theories emphasizes their distinctive approaches in the prolonged debate about the nature of emotion. However, despite their disagreements, these theories share a common assumption on the essential link between affective experience and expression. While the theories debate the nature of the signal, they all assume that a diagnostic association can be found between specific affective states and unique patterns of facial (Ekman, 1992) and vocal (Scherer, 1986) expressive outputs. Thereby, these theories assert a direct, causal link between the emotional experience and its consequent manifestation in the face and voice.

Consider the case of a two people experiencing emotions, Dan who just discovered that he won the lottery and Jack who just discovered that his life savings were lost in a stock market crash. All three models of emotions would agree that the expressions of emotion would be different in the positive vs. negative scenarios. While the cause may be different basic emotions, different appraisal checks for intrinsic pleasantness, or different dimensions of valence, all models essentially predict that these two distinct emotional experiences would yield very different expressions, each diagnostic for the valence of its occurring situation.

Shared Limitations in Emotion Perception Research

Essentially all models of emotion make critical use of emotional expression stimuli in testing predictions from their theory. However, although the hallmark of emotional expressions is their spontaneous expressive nature, the prevalent approach in contemporary studies of emotional expressions relies on *simulated* expressions of actors. In the preparation of such sets, professional or lay actors are instructed to portray various types of emotional expressions based on specific face movements, emotion labels, and/or acting typical affective scenarios. The simulated expressions are then presented to participants who either categorize the expressed emotion or rate its affective valence and arousal (for further discussion, see Scherer, 2003). Notwithstanding their advantages (highly homogeneous and consensual portrayals), simulated portrayals of emotion are cogently criticized for being overly simplified (Scherer, 1986) and stereotypical. Such expressions may fail to capture the complexity and variability of emotional displays in real life (Anikin & Lima, 2017; Fernández-Dols & Crivelli, 2013), and may thus inflate recognizability. In light of this criticism, there is a growing interest in pursuing more ecological and naturalistic emotional stimuli in emotion perception research. As next reviewed, this approach has generated important insights regarding the mechanisms of emotion perception in real life.

Naturalistic Emotional Expressions

Until recently, using naturalistic human expressions in research was methodologically challenging. Adequate documentation of people expressing emotions in natural situations was extremely scarce, often suffering from poor source quality in comparison to lab-created stimuli. However, the massive growth in digital media yielded exciting new opportunities to effectively integrate high-quality, real-life emotional expressions in contemporary research of emotion, with surprising results.

Affective Facial Expressions in Real-Life, Highly Intense Situations

Aviezer, Trope, and Todorov (2012a) used authentic photographs of professional tennis players, winning and losing points in a high-stakes match. Participants in that study were asked to judge the expressive reactions of winners and losers, by rating the affective valence of the facial expressions. Contrary to lay intuition and predictions of classic models, participants failed to discriminate between winners and losers based on their facial expressions alone, as both winning and losing faces were judged as conveying negative valence. These findings suggest that facial expressions alone may not be diagnostic for valence, at least during highly intense situations.

Utilizing facial expressions of athletes during professional sports events provides an exciting opportunity to test facial expressions of extreme emotions that normally could not be elicited in lab conditions. However, professional sports events may differ from day-to-day social interactions. The presence of a massive crowd and the intense physical activity during sports competitions may affect the facial displays of athletes. To meet this criticism, researchers can utilize another promising source of naturalistic expressions—social media. The ubiquity of social media has led millions of worldwide users to habitually document and share their everyday experiences via social networks like YouTube, Instagram, and Facebook. Some of these experiences involve intense emotional reactions, expressed spontaneously during real-life social interactions. For example, Wenzler, Levine, van Dick, Oertel-Knöchel, and Aviezer (2016) used authentic facial reactions to intense real-life positive (e.g., reactions of family members to homecoming soldier reunions) and negative (e.g., reactions to witnessing terror attacks) situations and tested their diagnosticity for valence. Consistent with previous findings from naturalistic facial displays of athletes (Aviezer et al., 2012a), participants failed to differentiate between real-life positive and negative emotional expressions, and both expression categories were judged as conveying negative affect.

Critically, past research relied mostly on static images that portray a single peak frame of the emotional display. This methodology may pose serious limitations, especially when testing naturalistic emotional expressions. In real life, emotional expressions are intrinsically dynamic and convey rich emotional information that unfolds and changes over time (Ambadar, Schooler, & Cohn, 2005). If dynamic expressions convey unique emotional information that is missing when using static or multi-static expressions, it may be essential to test the diagnosticity of naturalistic emotional expressions using real-life dynamic stimuli.

To this end, Israelashvili, Hassin, and Aviezer (2019) presented participants with real-life videos of people reacting to extreme positive events (family member's joyful reunions with homecoming soldiers). Videos were edited at various durations (5–20 s) to include fairly extensive portrayals of the dynamic unfolding of the face. Consistent with previous findings on static naturalistic stimuli (Aviezer et al., 2012a; Wenzler et al., 2016), dynamic real-life facial expressions in isolation were non-diagnostic for valence and were rated as conveying negative affect, irrespective of video duration.

Affective Vocal Expressions in Real-Life, Highly Intense Situations

While previous studies focused mostly on facial expressions, vocal expressions also serve as a rich source of affective information (Russell et al., 2003). Indeed, past studies have demonstrated perceptual and acoustic differences between vocal expressions of distinct emotion categories (Keltner, Tracy, Sauter, Cordaro, & McNeil, 2016). However, as noted earlier, previous work on affective vocalizations relied heavily on *simulated* expressions of actors, and only a handful of studies so far tested the perception of naturalistic vocal expressions in humans.

In setting out to test the diagnosticity of naturalistic human vocalizations, Atias et al. (2019) utilized intense vocal expressions evoked in a variety of real-life, positive and negative affective situations. For example, they used vocal reactions during a joyful reunion with a homecoming soldier vs. reactions to a terrifying encounter with an invader to one's home. Participants were asked to rate the perceived valence and intensity of each vocalization. Consistent with previous findings on naturalistic facial expressions, participants failed to differentiate the valence of these vocalizations—they were all judged as conveying negative affect. Furthermore, perceived intensity was strongly and negatively correlated with perceived valence, such that the more intense vocal reactions were, the more confusable they were to perceivers.

To further examine the role of intensity in driving perceptual ambiguity, Atias et al. (2019) used vocal reactions of real-life lottery winners. This unique set of vocalizations was taken from a subscription program of the Israeli National Lottery in which people sign up for a weekly lottery and can win various amounts of monetary prizes (ranging from ~$15K to ~$500K). Winners are called (and recorded) by an official lottery representative who notifies them of their winning sum. This allowed the researchers to examine how a real-life manipulation of hedonic intensity (i.e., winning more money) changes the perceived affect of the vocal expression. As predicted, vocal reactions to lower-sum wins were rated as positive, but reactions to higher-sum wins (from ~$60K and up) actually sounded negative.

In another study, Anikin and Persson (2017) introduced an exceptionally large corpus of 260 nonverbal vocalization obtained from online documentations of real-life affective situations. Affective situations in this corpus were selected to correspond to nine emotional categories previously explored using simulated vocal expressions of actors (i.e., amusement, anger, disgust, effort, fear, joy, pain, pleasure, and sadness). Accuracy rates and confusion patterns in classifying the emotional categories of these vocalizations were revealing. For example, cross valence errors, a virtually nonexistent confusability in posed vocalizations (Belin et al., 2008), occurred quite frequently such that most of the joy reactions were classified as fear. Additionally, the distinction between basic emotions such as fear and anger was poor, with roughly half the angry vocalizations classified as fearful. Interestingly, while participants agreed on naming the call types (e.g., laugh, scream, cry), they differed vastly in how these call types map onto emotions (e.g., amusement, fear, sadness, respectively; Anikin, Bååth, & Persson, 2018). As noted by Anikin et al. "… when a participant classified a sound as a scream, the perceived emotion varied widely and included quite distinct contexts, such as fear, pain, delight, surprise, etc.".

While intriguing, the findings from both real-life faces and voices pose a puzzle: If naturalistic emotional expressions do not necessarily convey valid diagnostic information for affect, what is the mechanism that facilitates emotion communication in everyday life? We next suggest a plausible answer to this question, demonstrating the role of contextual information in the perception of emotional expression.

Naturalistic Emotional Expressions in Context

In the previous section, we discussed the ambiguity of naturalistic emotional expressions, especially during highly intense situations. Importantly, this ambiguity is most evident in conditions where emotional expressions are presented in isolation (i.e., without context). However, in real life, emotional expressions are rarely manifested in total isolation, rather, they are typically embedded in rich context (i.e., any source of information external to the emotional expression itself; Hassin, Aviezer, & Bentin, 2013).

Hence, to gain a better understanding of emotion communication in real life, recent studies started exploring the role of context in the perception of naturalistic

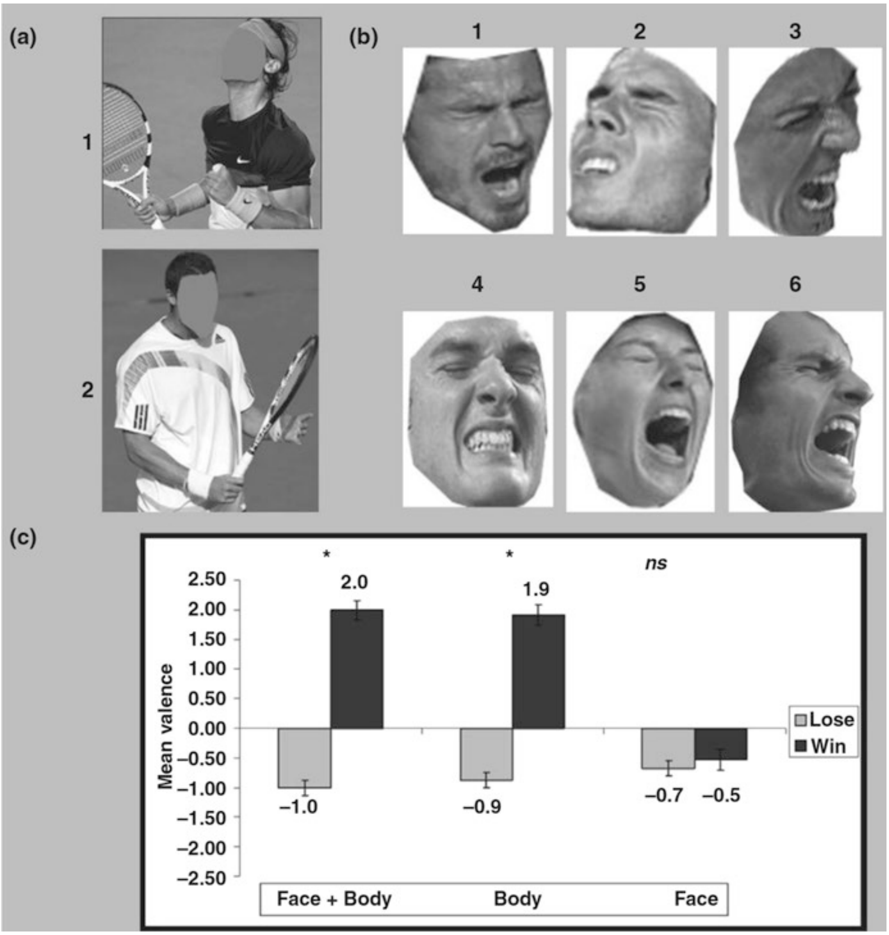


Fig. 1 (**a**) Characteristic body posture of (1) winners and (2) losers. (**b**) Isolated facial expressions of winners and losers in tennis (1, 4, 6—losing point; 2, 3, 5—winning point). (**c**) Mean valence ratings when images were presented in face only, body only, and face with body formats. (All photos in this figure courtesy of a.s.a.p. Creative/Reuters. Adapted from *Science* with permission)

emotional expressions. Aviezer et al. (2012a) demonstrated that while real-life intense faces of athletes were highly ambiguous, they were well recognized when perceived with their gesturing bodies (Fig. 1). Furthermore, they demonstrated the high susceptibility of naturalistic facial expressions to contextual influence by combining winning and losing facial expressions of athletes with congruent and incongruent contextual body postures. As predicted, valence ratings of facial expressions shifted categorically (from positive to negative and vice versa) depending on the valence conveyed by the accompanied body postures, even when participants explicitly relied on the face when making their judgments.

Similar patterns of contextual influence were also demonstrated when dynamic facial expressions were embedded in authentic videos of real-life events (Israelashvili et al., 2019). In a recent striking demonstration of the importance of context in emotion perception, researchers entirely masked characters' faces and bodies in silent videos. Nevertheless, viewers inferred the affect of the invisible characters successfully and in high agreement based solely on visual context (Chen & Whitney, 2019).

Atias et al. (2019) tested the influence of visual context on the perception of emotional vocalizations by combining posed and real-life emotional vocalizations with real-life videos of people reacting to extreme positive events (e.g., homecoming reunions with family members) and negative events (e.g., encounters with invaders to one's home). Participants were asked to rate the valence of each vocalization regardless of the visual context. Remarkably, valence ratings of the vocalizations shifted categorically from positive to negative and vice versa, according to the accompanying video context. Furthermore, real-life vocalizations were found more susceptible to the effect of context than posed vocalizations (see Fig. 2).

This dramatic susceptibility of naturalistic emotional expressions to context highlights a critical characteristic of the emotion perception process that is typically ignored in classic research. Traditionally, emotion perception was explored by presenting isolated affective cues (e.g., facial expressions, speech prosody) and testing their diagnosticity in controlled settings. By contrast, an alternative approach in interpreting the contextual malleability of emotional expressions suggests that emotion perception is not determined by the mere extraction of meaningful affective information from isolated cues. Rather, emotion perception can be conceptualized

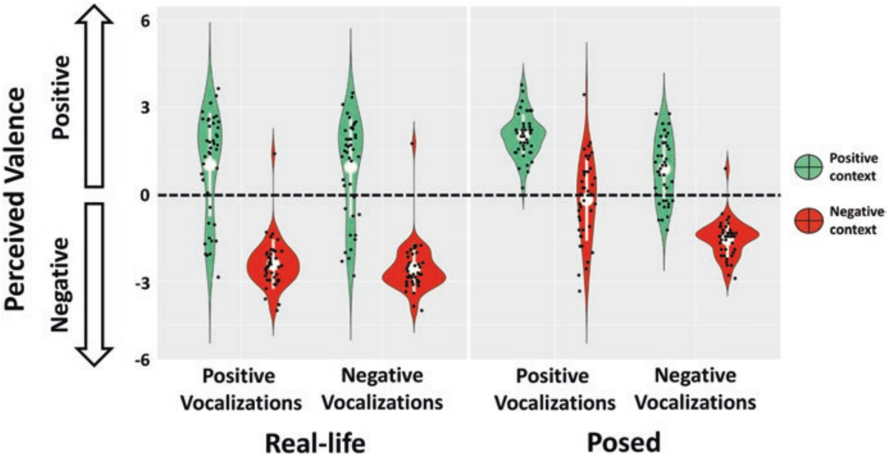


Fig. 2 Violin plots of perceived valence of real-life and posed, positive and negative vocalizations as a function of positive and negative context (central white circles and vertical white bars reflect mean and SD). Context dramatically influenced the perceived valence of vocalizations such that the same vocalizations sounded positive or negative when paired with differently valenced visual context, an effect more robust for the real-life stimuli than for the posed stereotypical vocalizations. (Adapted from *Journal of Experimental Psychology: General* with permission)

as a holistic process by which various sources of information are integrated into a unique, gestalt-like percept that cannot be attributed solely to one source or another (Aviezer, Ensenberg, & Hassin, 2017; Aviezer, Trope, & Todorov, 2012b).

This conceptualization has a long tradition in psychological science and is supported by studies demonstrating cross-modal integration of sensory cues. The recurrent evidence from these studies shows that the integration of multisensory cues facilitates perception (de Gelder & Vroomen, 2000) and often generates completely novel percepts that could not be accounted solely by each source alone (McGurk & MacDonald, 1976; Shams, Kamitani, & Shimojo, 2002). The robustness of multisensory integration and its impact on perception is also supported by neuroimaging studies that show convergent evidence for cross-modal processing not only in multisensory cortical and subcortical areas, but also in primary sensory regions that were once believed to be modality-specific (Amedi, von Kriegstein, van Atteveldt, Beauchamp, & Naumer, 2005; Kayser & Logothetis, 2007; Senkowski, Schneider, Foxe, & Engel, 2008). This suggests that in order to better model emotion perception in real life, it is essential to employ research paradigms that facilitate the integration of affective cues from different sources of information, rather than focus on isolated expressions. To this end, testing the perception of naturalistic emotional expression in context may provide a promising direction towards delineating the mechanisms that guide emotion perception in real life.

Empathy and Emotion Perception

Let us return to the hurt child with which we opened our chapter. Classic models of emotion perception inform us that specific expressive cues in the child's face and voice are decoded by the father, informing him about the feeling of the child and guiding the empathic response. While intuitive, the conclusion of our brief review on real-life emotion perception suggests a more cautious approach, because the expressive cues, per se, may be far more ambiguous than previously assumed. As such, the importance of recognizing isolated emotional expressions in empathy (e.g., Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001; Besel & Yuille, 2010; Clark et al., 2008) may have been overstated, while the importance of contextual information may have been downplayed.

Empathic Accuracy and the Perception of Emotion

As noted, a central paradigm for assessing the ability of perceivers' to identify the thoughts and feelings of social targets is the empathic accuracy paradigm (Ickes, 1993; Zaki, Weber, Bolger, & Ochsner, 2009). Empathic accuracy can be quantified using an affect-rating paradigm that begins with targets describing emotional autobiographical events while being videotaped by the experimenter. Then, targets watch their own videos and provide ground-truth judgments of how they felt while discussing these events. In a subsequent session, naïve perceivers watch targets' videos and provide a similar judgment of the affect they believe the targets experienced in the video. Different paradigms have been offered for measuring the concordance between the target and the perceiver. For example, in the work of Gesn and Ickes (1999), targets undergoing therapy viewed a video of their session while verbally reporting their thoughts and feelings at various points. In the work of Zaki, Bolger, and Ochsner (2009), targets told an emotional story and then viewed a video of themselves while reporting their affect using a nonverbal continuous valence-rating scale. Empathic accuracy was then calculated as the time-course correlation between targets' self-reported affect and perceivers' judgments of targets' affect (Ickes et al., 1990; Levenson & Ruef, 1992; Zaki et al., 2008).

Previous work had shown that empathic accuracy is strongly dependent on the targets' expressivity, suggesting that the affective cues and emotional expressions conveyed by targets during social interactions are central to empathic accuracy (Zaki et al., 2008). However, the link between expressions and empathic accuracy is not straightforward. For example, Gesn and Ickes (1999) tested empathic accuracy when perceivers were exposed to three different types of information channels: original full videos, audio only, and videos plus electronically filtered audio (designed to make semantic information unintelligible while preserving speech prosody). Results showed that empathic accuracy was dramatically higher in the two conditions that contained verbal information (the full video and auditory only conditions) in comparison to the condition that omitted verbal information (the video plus filtered audio condition). This led to the conclusion that while nonverbal cues may provide some contribution to empathic accuracy, auditory verbal information appears to be the most diagnostic channel for empathic accuracy.

Similar results indicating the inherent ambiguity of nonverbal information were obtained using the continuous affective valence rating task (Zaki, Bolger, et al., 2009). In that study, perceivers were moderately accurate about target affect ($r = 0.47$) when rating the video with audio (which included verbal information). However, accuracy dropped dramatically ($r = 0.21$) when the video was presented without the audio and participants were forced to rely on nonverbal information alone.

In fact, there are reasons to believe that even this very modest correlation is inflated compared to real-life affect recognition. First, the expressive individuals were telling a story to a camera, a conversational method that increases the use of visual prosody, a linguistic strategy involving head and face musculature movements that aid speech communication (Graf, Cosatto, Strom, & Huang, 2002; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). Real-life emotional expressions may be less controlled and more ambiguous. Second, the emotional intensity felt by people telling stories is arguably weaker than of people experiencing the events in real time. Given our reviewed findings showing that intense emotional expressions may be non-diagnostic, relying on the isolated faces or voices would lead to blatant errors in empathic accuracy.

Empathic Accuracy and Contextualized Emotion Perception

While isolated nonverbal information is largely non-diagnostic in empathic accuracy paradigms, it may still have an important role in contributing to contextualized empathic accuracy. For example, in a study which contrasted spoken words vs. words in a transcript (with no vocalizations or intonation available), the former displayed higher empathic accuracy (Hall & Schmid Mast, 2007). This suggests that verbal cues may be a dominant source of information in producing empathic accuracy, but only when integrated with nonverbal vocal information as well. Similarly, in the study by Zaki, Bolger, et al. (2009), empathic accuracy was higher to full videos with audio, than to audio alone. To us these findings suggest bi-directional, dynamic integration effects by which ambiguous expressions are disambiguated by semantic context, and in return, serve as feedback, accentuating and emphasizing the semantic information itself.

This critical role of source integration in deriving empathic accuracy is in good accordance with the aforementioned context effects found in emotion perception research, suggesting that both processes may rely on the holistic evaluation of targets' expressive behavior (Ickes, 2001). Consequently, empathic accuracy, like emotion perception, may not be determined by a specific dominant source of information. Rather, it likely relies on flexible integration of affective cues from various channels into a unique, contextually coherent percept that facilitates interpersonal interaction. From this perspective, the extensive focus on expressive targets as the principal informational basis for empathic accuracy may pose some serious limitations.

Emotion perception as well as empathic accuracy may not be about recognizing isolated cues, but rather, about successfully integrating cues with context. In the field of emotion perception, recent work on individual differences suggests that perceivers consistently vary in their tendencies to integrate contextual information when making affect judgments of facial expressions. Some perceivers tend to rely solely on targets' facial expressions, while ignoring the information derived from the context. Others show the opposite pattern and tend to be very easily swayed by contextual information, with minimal influence of the facial expression (Aviezer et al., 2017).

We suggest that individual differences in empathic accuracy could also be conceptualized in terms of similar integration processes. For example, insufficient integration of affective cues may lead perceivers to focus on limited sources of information during interpersonal interactions, which may impede empathic accuracy. This may suggest a plausible mechanism for the poor performance of individuals with autism in empathic accuracy tasks, related to their detail-focused processing style and general failure to extract a gestalt-like percept (Behrmann, Thomas, & Humphreys, 2006; Demurie, De Corel, & Roeyers, 2011).

By contrast, an excessive integration of irrelevant contextual information during interpersonal interactions may distort and bias perceivers' inferences of targets' actual thoughts and feelings and impede empathic accuracy. For example, individuals with social anxiety may exhibit biased emotion perception due to the activation of dysfunctional interpersonal beliefs during social interactions (Silvia, Allan, Beauchamp, Maschauer, & Workman, 2006). One way to pursue this direction and still maintain high ecological validity is to utilize naturalistic stimuli of extreme emotions. As demonstrated, extreme emotional expressions are inherently ambiguous, and adequate judgments of these expressions rely heavily on the integration of contextual information. Therefore, empathic accuracy of extreme emotions may vary considerably as a function of perceivers' integration abilities. These suggestive mechanisms for individual differences in empathic accuracy have yet to be empirically tested.

Coda

Research on emotion perception has progressed dramatically in recent years with the development of innovative research paradigms that utilize naturalistic and contextualized emotional expressions. The converging evidence from these studies suggests that naturalistic emotional expressions are far more ambiguous and contextually malleable than previously assumed, especially in highly intense situations. The immense role of context in the perception of naturalistic emotional expression has far-reaching implications not only for the field of emotion perception but also for the closely related research on empathic accuracy. Contextual integration is likely involved in our ability to feel the emotions of other individuals, not as a phenomenon at the fringe of empathic accuracy, but rather at its very core.

Acknowledgments This work was supported by an Israel Science Foundation [ISF#259/18] grant to Hillel Aviezer.

References

Ambadar, Z., Schooler, J. W., & Cohn, J. F. (2005). Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions. *Psychological Science, 16*(5), 403–410. https://doi.org/10.1111/j.0956-7976.2005.01548.x

Amedi, A., von Kriegstein, K., van Atteveldt, N. M., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research, 166*(3), 559–571. https://doi.org/10.1007/s00221-005-2396-5

Anikin, A., Bååth, R., & Persson, T. (2018). Human non-linguistic vocal repertoire: Call types and their meaning. *Journal of Nonverbal Behavior, 42*(1), 53–80. https://doi.org/10.1007/s10919-017-0267-y

Anikin, A., & Lima, C. F. (2017). Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations. *Quarterly Journal of Experimental Psychology, 17470218*(2016), 1. https://doi.org/10.1080/17470218.2016.1270976

Anikin, A., & Persson, T. (2017). Nonlinguistic vocalizations from online amateur videos for emotion research: A validated corpus. *Behavior Research Methods, 49*(2), 758–771. https://doi.org/10.3758/s13428-016-0736-y

Atias, D., Todorov, A., Liraz, S., Eidinger, A., Dror, I., Maymon, Y., & Aviezer, H. (2019). Loud and unclear: Intense real-life vocalizations during affective situations are perceptually ambiguous and contextually malleable. *Journal of Experimental Psychology: General, 148*(10), 1842. https://doi.org/10.1037/xge0000535

Aviezer, H., Ensenberg, N., & Hassin, R. R. (2017). The inherently contextualized nature of facial emotion perception. *Emotion, 17*, 47–54. https://doi.org/10.1016/j.copsyc.2017.06.006

Aviezer, H., Trope, Y., & Todorov, A. (2012a). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science, 338*(6111), 1225–1229. https://doi.org/10.1126/science.1224313

Aviezer, H., Trope, Y., & Todorov, A. (2012b). Holistic person processing: Faces with bodies tell the whole story. *Journal of Personality and Social Psychology, 103*(1), 20–37. https://doi.org/10.1037/a0027411

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614–636. https://doi.org/10.1037/0022-3514.70.3.614

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines, 42*(2), 241–251.

Behrmann, M., Thomas, C., & Humphreys, K. (2006). Seeing it differently: Visual processing in autism. *Trends in Cognitive Sciences, 10*(6), 258–264. https://doi.org/10.1016/j.tics.2006.05.001

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods, 40*(2), 531–539. https://doi.org/10.3758/BRM.40.2.531

Besel, L. D. S., & Yuille, J. C. (2010). Individual differences in empathy: The role of facial expression recognition. *Personality and Individual Differences, 49*(2), 107–112. https://doi.org/10.1016/j.paid.2010.03.013

Chen, Z., & Whitney, D. (2019). Tracking the affective state of unseen persons. *Proceedings of the National Academy of Sciences, 116*(15), 7559. https://doi.org/10.1073/pnas.1812250116

Clark, T. F., Winkielman, P., & McIntosh, D. N. (2008). Autism and the extraction of emotion from briefly presented facial expressions: Stumbling at the first step of empathy. *Emotion, 8*(6), 803–809. https://doi.org/10.1037/a0014124

Darwin, C. (1872). *Expression of the emotions in man and animals*. London: Albemarle.

de Gelder, B. (2006). Towards the neurobiology of emotional body language. *Nature Reviews Neuroscience, 7*(3), 242–249. https://doi.org/10.1038/nrn1872

de Gelder, B. (2009). Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 364*(1535), 3475–3484. https://doi.org/10.1098/rstb.2009.0190

de Gelder, B., & Van den Stock, J. (2011). The bodily expressive action stimulus test (BEAST). Construction and validation of a stimulus basis for measuring perception of whole body expression of emotions. *Frontiers in Psychology, 2*, 181. https://doi.org/10.3389/fpsyg.2011.00181

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition & Emotion, 14*(3), 289–311. https://doi.org/10.1080/026999300378824

Demurie, E., De Corel, M., & Roeyers, H. (2011). Empathic accuracy in adolescents with autism spectrum disorders and adolescents with attention-deficit/hyperactivity disorder. *Research in Autism Spectrum Disorders, 5*(1), 126–134. https://doi.org/10.1016/j.rasd.2010.03.002

Ekman, P. (1976). *Pictures of facial affect*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*, 384.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: Guidelines for research and an integration of findings*. Oxford: Pergamon Press.

Ekman, P., & Rosenberg, E. (Eds.). (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195179644.001.0001

Fernández-Dols, J.-M., & Crivelli, C. (2013). Emotion and expression: Naturalistic studies. *Emotion Review, 5*(1), 24–29. https://doi.org/10.1177/1754073912457229

Fridlund, A. J. (1994). *Human facial expression: An evolutionary view*. San Diego, CA: Academic Press.

Gendron, M., Lindquist, K. A., Barsalou, L., & Barrett, L. F. (2012). Emotion words shape emotion percepts. *Emotion, 12*(2), 314–325. https://doi.org/10.1037/a0026007

Gesn, P. R., & Ickes, W. (1999). The development of meaning contexts for empathic accuracy: Channel and sequence effects. *Journal of Personality and Social Psychology, 77*(4), 746–761. https://doi.org/10.1037/0022-3514.77.4.746

Graf, H. P., Cosatto, E., Strom, V., & Huang, F. J. (2002). Visual prosody: Facial movements accompanying speech. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition* (pp. 396–401). Washington, DC: IEEE. https://doi.org/10.1109/AFGR.2002.1004186

Hall, J. A., & Schmid Mast, M. (2007). Sources of accuracy in the empathic accuracy paradigm. *Emotion, 7*(2), 438–446. https://doi.org/10.1037/1528-3542.7.2.438

Hassin, R. R., Aviezer, H., & Bentin, S. (2013). Inherently ambiguous: Facial expressions of emotions, in context. *Emotion Review, 5*(1), 60–65.

Hawk, S. T., Van der Schalk, J., & Fischer, A. H. (2008). Moving faces, looking places: The Amsterdam dynamic facial expressions set (ADFES). In *12th European Conference on Facial Expressions, Geneva, Switzerland*.

Ickes, W. (1993). Empathic accuracy. *Journal of Personality, 61*(4), 587–610. https://doi.org/10.1111/j.1467-6494.1993.tb00783.x

Ickes, W. (2001). *Measuring empathic accuracy. Interpersonal sensitivity: Theory and measurement* (pp. 219–241). Mahwah, NJ: Lawrence Erlbaum Associates.

Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). Naturalistic social cognition: Empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology, 59*(4), 730–742. https://doi.org/10.1037/0022-3514.59.4.730

Israelashvili, J., Hassin, R. R., & Aviezer, H. (2019). When emotions run high: A critical role for context in the unfolding of dynamic, real-life facial affect. *Emotion, 19*(3), 558–562. https://doi.org/10.1037/emo0000441

Kayser, C., & Logothetis, N. K. (2007). Do early sensory cortices integrate cross-modal information? *Brain Structure and Function, 212*(2), 121–132. https://doi.org/10.1007/s00429-007-0154-0

Keltner, D., Sauter, D., Tracy, J., & Cowen, A. (2019). Emotional expression: Advances in basic emotion theory. *Journal of Nonverbal Behavior, 43*(2), 133–160. https://doi.org/10.1007/s10919-019-00293-3

Keltner, D., Tracy, J., Sauter, D. A., Cordaro, D. C., & McNeil, G. (2016). Expression of emotion. In *Handbook of emotions* (pp. 467–482). New York, NY: Guilford Press.

Levenson, R. W., & Ruef, A. M. (1992). Empathy: A physiological substrate. *Journal of Personality and Social Psychology, 63*(2), 234–246. https://doi.org/10.1037/0022-3514.63.2.234

Lindquist, K. A., & Barrett, L. F. (2008). Constructing emotion: The experience of fear as a conceptual act. *Psychological Science, 19*(9), 898–903. https://doi.org/10.1111/j.1467-9280.2008.02174.x

Lindquist, K. A., & Gendron, M. (2013). What's in a word? Language constructs emotion perception. *Emotion Review, 5*(1), 66–71. https://doi.org/10.1177/1754073912451351

Marsh, A. A., Kozak, M. N., & Ambady, N. (2007). Accurate identification of fear facial expressions predicts prosocial behavior. *Emotion, 7*(2), 239–251. https://doi.org/10.1037/1528-3542.7.2.239

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746–748. https://doi.org/10.1038/264746a0

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science, 15*(2), 133–137. https://doi.org/10.1111/j.0963-7214.2004.01502010.x

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology, 17*(3), 715–734. https://doi.org/10.1017/S0954579405050340

Roseman, I. J. (1984). Cognitive determinants of emotion: A structural theory. *Review of Personality & Social Psychology, 5*, 11–36.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Russell, J. A. (1997). Reading emotions from and into faces: Resurrecting a dimensional-contextual perspective. In *The psychology of facial expression* (pp. 295–320). Cambridge: Cambridge University Press.

Russell, J. A., Bachorowski, J.-A., & Fernández-Dols, J.-M. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology, 54*(1), 329–349. https://doi.org/10.1146/annurev.psych.54.101601.145102

Russell, J. A., & Bullock, M. (1986). Fuzzy concepts and the perception of emotion in facial expressions. *Social Cognition, 4*(3), 309–341. https://doi.org/10.1521/soco.1986.4.3.309

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences, 107*(6), 2408–2412. https://doi.org/10.1073/pnas.0908239106

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2015). Emotional vocalizations are recognized across cultures regardless of the valence of distractors. *Psychological Science, 26*(3), 354–356. https://doi.org/10.1177/0956797614560771

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer & D. D. Clarke (Eds.), *Approaches to emotion*. Hillsdale, NJ: Erlbaum. Retrieved from http://www.nottingham.ac.uk/is/gateway/esl/local/pdf/0898594065(293-317).pdf

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99*(2), 143–165. https://doi.org/10.1037/0033-2909.99.2.143

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*(1), 227–256. https://doi.org/10.1016/S0167-6393(02)00084-5

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion, 23*(7), 1307–1351. https://doi.org/10.1080/02699930902928969

Scherer, K. R., & Moors, A. (2019). The emotion process: Event appraisal and component differentiation. *Annual Review of Psychology, 70*(1), 719–745. https://doi.org/10.1146/annurev-psych-122216-011854

Scherer, K. R., Mortillaro, M., & Mehu, M. (2017). Facial expression is driven by appraisal and generates appraisal inference. In *The science of facial expression* (pp. 353–373). New York, NY: Oxford University Press.

Schlosberg, H. (1941). A scale for the judgment of facial expressions. *Journal of Experimental Psychology, 29*(6), 497–510. https://doi.org/10.1037/h0061489

Schlosberg, H. (1952). The description of facial expressions in terms of two dimensions. *Journal of Experimental Psychology, 44*(4), 229–237. https://doi.org/10.1037/h0055778

Senkowski, D., Schneider, T. R., Foxe, J. J., & Engel, A. K. (2008). Crossmodal binding through neural coherence: Implications for multisensory processing. *Trends in Neurosciences, 31*(8), 401–409. https://doi.org/10.1016/j.tins.2008.05.002

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Multisensory Proceedings, 14*(1), 147–152. https://doi.org/10.1016/S0926-6410(02)00069-1

Silvia, P. J., Allan, W. D., Beauchamp, D. L., Maschauer, E. L., & Workman, J. O. (2006). Biased recognition of happy facial expressions in social anxiety. *Journal of Social and Clinical Psychology, 25*(6), 585–602. https://doi.org/10.1521/jscp.2006.25.6.585

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology, 48*(4), 813–838. https://doi.org/10.1037/0022-3514.48.4.813

Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005). Transmitting and decoding facial expressions. *Psychological Science, 16*(3), 184–189. https://doi.org/10.1111/j.0956-7976.2005.00801.x

Tomkins, S. S. (1962). *Affect, imagery, consciousness: Vol. 1. The positive affects*. New York, NY: Springer. https://doi.org/10.1037/14351-000

Tomkins, S. S. (1963). *Affect, imagery, consciousness: Vol. 2. The negative affects*. New York, NY: Springer.

Tracy, J. L., & Matsumoto, D. (2008). The spontaneous expression of pride and shame: Evidence for biologically innate nonverbal displays. *Proceedings of the National Academy of Sciences, 105*(33), 11655. https://doi.org/10.1073/pnas.0802686105

Tracy, J. L., & Robins, R. W. (2004). Show your pride: Evidence for a discrete emotion expression. *Psychological Science, 15*(3), 194–197. https://doi.org/10.1111/j.0956-7976.2004.01503008.x

Tracy, J. L., & Robins, R. W. (2008). The nonverbal expression of pride: Evidence for cross-cultural recognition. *Journal of Personality and Social Psychology, 94*(3), 516.

Wenzler, S., Levine, S., van Dick, R., Oertel-Knöchel, V., & Aviezer, H. (2016). Beyond pleasure and pain: Facial expression ambiguity in adults and children during intense situations. *Emotion, 16*(6), 807–814. https://doi.org/10.1037/emo0000185

Woodworth, R. S., & Schlosberg, H. (1938). *Experimental psychology*. New York, NY: Henry Holt and Company.

Zaki, J., Weber, J., Bolger, N., & Ochsner, K. (2009). The neural bases of empathic accuracy. *Proceedings of the National Academy of Sciences, 106*(27), 11382–11387. https://doi.org/10.1073/pnas.0902666106

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological Science, 19*(4), 399–404. https://doi.org/10.1111/j.1467-9280.2008.02099.x

Zaki, J., Bolger, N., & Ochsner, K. (2009). Unpacking the informational bases of empathic accuracy. *Emotion, 9*(4), 478–487. https://doi.org/10.1037/a0016551

Zaki, J., & Ochsner, K. (2011). The cognitive neuroscience of sharing and understanding others' emotions. In J. Decety (Ed.), *Empathy: From bench to bedside*. Cambridge, MA: MIT Press.

Flexible Social Cognition: A Context-Dependent Failure to Mentalize

Melissa Jhurry and Lasana T. Harris

People spontaneously get inside the heads of others (mentalize) to explain and predict that person's (social target's) behavior. As such, mentalizing facilitates other forms of social cognition such as empathy, perspective-taking, and emotional resonance, and is a fundamental component of social interactions. Mentalizing even occurs to entities that do not possess mind—a processes termed anthropomorphism—underlining its pervasive nature. A failure to mentalize is therefore striking given its usefulness and pervasive nature. In this chapter, we will explore the boundary conditions of mentalizing by focusing on a variety of social contexts that promote flexible social cognition, focusing on failures to mentalize rather than anthropomorphism. We will also describe the brain networks and regions that govern the modulation of mentalizing, and explore the role other cognitive and affective processes play in promoting or inhibiting mentalizing.

Social Cognition Brain Network

Mentalizing is necessary to enable social cognition. Previous neuroimaging studies within the field of social neuroscience have reached a consensus on the network of brain regions that underlie social cognition, a social cognition brain network (SCBN) that includes the medial prefrontal cortex (MPFC), the superior temporal sulcus (STS), the precuneus, the posterior cingulate cortex (PCC), the tempoparietal junction (TPJ), and the temporal poles (Amodio & Frith, 2006; Mars et al., 2012; Van Overwalle, 2009). There are also other brain regions that are involved in specific social cognitive processes beyond those mentioned, and all the brain regions within the SCBN also play a role in other non-social cognitive processes. The

M. Jhurry · L. T. Harris (✉)
Department of Experimental Psychology, University College London, London, UK
e-mail: melissa.jhurry.18@ucl.ac.uk; lasana.harris@ucl.ac.uk

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_10

complexity of the SCBN mirrors the complexity of the social cognitive mechanisms that integrate various types of statistical information to produce a variety of outputs. Therefore, social cognition cannot be identified by the activity of a single brain region; rather, it is the simultaneous activity of the entire SCBN that can act as an index of social cognition engagement.

A brain region initially implicated in mentalizing ability and social cognition is the MPFC (Amodio & Frith, 2006), specifically associated with making inferences about a social target's enduring disposition (Harris, Todorov, & Fiske, 2005). On the other hand, the TPJ appears to be especially involved in inferring short-term mental states, such as current beliefs, goals, and emotions (Van Overwalle, 2009). The STS and temporal poles are also reliably activated by tasks that require mentalizing ability (Gallagher & Frith, 2003). The STS has been implicated in the detection of biological motion and appears to be activated by the perception of movement or the implied movement of an agentic other (Allison, Puce, & McCarthy, 2000; Frith & Frith, 2001; Gallagher & Frith, 2003). The STS has also been implicated in visual perspective-taking and the observation of eye movements (Frith & Frith, 2006). Many studies have found evidence of temporal pole activation during mentalizing tasks (Fletcher et al., 1995; Gallagher & Frith, 2003; Olson, Plotzker, & Ezzyat, 2007). The temporal poles have also been implicated in the retrieval of autobiographical episodic memory and personal semantic memory (Gallagher & Frith, 2003; Wiggs, Weisberg, & Martin, 1998), which includes socially relevant processes such as recognizing familiar faces and voices. In addition, the temporal poles have been associated with the processing of emotions (Decety & Jackson, 2006). Previous research suggests that the functions of the temporal poles extend beyond the recognition of sensory stimuli since they also link emotional responses to stimuli (Olson et al., 2007). The temporal poles have also been implicated in third-person perspective-taking in emotional contexts (Ruby & Decety, 2004). Finally, the precuneus and PCC have also been associated with multimodal first- and third-person perspective-taking and impression formation (Schilbach, Eickhoff, Rotarska-Jagiela, Fink, & Vogeley, 2008; Schiller, Freeman, Mitchell, Uleman, & Phelps, 2009), as well as the retrieval of episodic memories (Cavanna & Trimble, 2006).

Default Mode Network

While evidence for the SCBN has been established from studies utilizing social cognition tasks, a similar brain activation pattern has been observed in studies investigating the brain's default resting state; when the mind is not engaged in any specific cognitive task, the person is mind-wandering. The brain regions that underlie the default resting state, referred to as the default mode network (DMN), overlap significantly with the SCBN (Mars et al., 2012; Throop et al., 2004). One explanation for this overlap is a functional overlap between social cognition and the resting state; social cognition may be our default psychological state (Mars et al., 2012; Schilbach et al., 2008). In the resting state, participants engage in mind-wandering, mostly thinking about their social situations, and thereby engaging the SCBN (D'argembeau et al., 2005; Schilbach et al., 2008; Schooler et al., 2011). Evidence for social cognition as the psychological default comes from previous resting state experiments that have found reduced activity in the DMN when performing non-social cognitive control tasks, whereas social cognition tasks did not trigger any significant change in DMN activation levels compared to baseline (Iacoboni et al., 2004).

Mentalizing Overview

As mentioned above, mentalizing allows the perceiver to explain the behavior of the social target and to predict that social target's future behavior. Perceivers can also use mentalizing to manage the social target's impression of the perceiver; mentalizing engages reputation or impression management concerns (Fiske & Taylor, 1991). Perceivers can make such mental state inferences about various psychological states of the social target, including current goals and emotions, which are temporary mental states, or more long-term dispositional states, such as personality (Frith & Frith, 2006).

Mentalizing enables the perceiver to infer the social target's intentions. Warmth and competence are two primary dimensions of person perception (Asch, 1946; Fiske, Cuddy, Glick, & Xu, 2002; Wojciszke, 1994). Perceived warmth provides information about whether the other person has positive or negative intentions, while perceived competence provides information about the capability of the other person to achieve their intended goal (Fiske et al., 2002). Mentalizing is necessary for inferring warmth since intentions are mental constructs, while mentalizing is sufficient for competence inferences because they can be assessed by observing behavior independent of mentalizing.

According to attribution theory in social psychology, perceivers can integrate various types of information about a social target; such as internal information derived from mentalizing, and external factors derived from social norms, to explain the behaviors of others. There are many models that attempt to explain such integration. According to Kelley's (1973) covariation model, perceivers integrate three types of statistical information to attribute a cause for a social target's behavior: consistency—how the social target has behaved in the past, consensus—how other social targets generally behave, and distinctiveness—how the social target tends to behave in specific situations. Such models that describe mentalizing as the integration of statistical information suggest that mentalizing is a very cognitive and rational psychological process.

Further support for the rational nature of mentalizing comes from the notion of teleological reasoning (Gergely, Nádasdy, Csibra, & Bíró, 1995). The landmark study investigated the ability of 12-month-old infants to take the "intentional stance": to infer the intentions of an agent from their goal-directed behaviors. In order to take the intentional stance, it is necessary to be able to evaluate what is the most rational action the agent could take to achieve their goal in a new context. Gergely et al. (1995) presented 12-month-old infants with a circle travelling across the screen to reach a second circle on the opposite side. First, the infants viewed the stimuli with a rectangle obstructing the path between the two circles. The first circle would then jump over the rectangle to get to the circle on the other side. Then the infants were shown the same stimuli but without the obstacle of the rectangle between the circles. The results suggested that the infants were able to attribute agency to the circle and infer the agent's goal of reaching the second circle (Gergely et al., 1995). In addition, the infants showed more surprise in the no-obstacle condition where the agent would jump in the middle of the screen mimicking its old action, than if it took the novel but more rational action of travelling in a straight line. This suggests that the infants are not only able to mentalize, but that mentalizing is inherently rational.

Mentalizing not only assumes rationality of action, but morality of the agent as well. Support for such assertions comes from studies where infants infer the moral character of "good" and "bad" agents. One such study conducted a series of experiments where 6- and 10-month-old infants were presented with a wooden display depicting a hill and two characters, in the form of wooden shapes with eyes, interacting with each other. In the first experiment, the infants were presented with one character—the Climber, who attempts to climb the hill multiple times unsuccessfully until a second character—the Helper, pushes the Climber up the hill (Condition 1). In Condition 2, a second character—the Hinderer—pushes the Climber back down the hill. Once the infants had viewed this interaction, the experimenter offered the infant a choice of either the Helper or the Hinderer. The infants preferred the Helper over the Hinderer (Hamlin, Wynn, & Bloom, 2007). These results suggest that infants are able to infer the intentions of the characters via their actions and make at least a rudimentary moral judgement on whether the character is good or bad. While 6-month-old infants were able to make an evaluation of the moral goodness of the characters, by the age of 10 months, infants seemed to have developed the ability to attribute their own attitudes to the Climber (Hamlin et al., 2007).

A second study found that as well as being able to infer moral character (in other words, whether the agent was pro- or anti-social), infants were also able to use this information as the criteria to license helpful or harmful behaviors towards an agent (Hamlin, Wynn, Bloom, & Mahajan, 2011). This study tested 5- and 8-month-old infants on their ability to carry out such complex moral evaluations. In this experiment, puppets were used to introduce a pro- or anti-social agent to the infants, as in the previous study. In the second phase of the experiment, the infants then viewed the second puppet (either pro- or anti-social) playing with a ball and dropping it. In one trial, a giver puppet picks up the dropped ball and returns it. In a second trial, a taker puppet takes the ball away. Both 5-month-old and 8-month-old infants preferred the giver in the pro-social condition (where the puppet playing with the ball had been pro-social in the first phase of the experiment). However, in the anti-social condition, the 8-month-old infants preferred the taker. Meanwhile, the 5-month-old infants preferred the giver in both pro- and anti-social conditions (Hamlin et al., 2011). Like the first study by Hamlin et al. (2007), the infants have the ability to distinguish between the “good” and “bad” agents; however, the older infants have developed the ability to make more complex moral judgements, in this case being able to license negative behaviors towards an anti-social other, while still showing a general preference for pro-social behaviors (Hamlin et al., 2011).

While these studies suggest that infants can use their mentalizing ability to facilitate moral evaluation, another study demonstrated that mentalizing ability also facilitates innate altruism (Warneken & Tomasello, 2006). In this study, experimenters observed 18-month-old infants who were in the presence of an adult experimenter struggling to achieve a series of goals. These included four different scenarios: “The adult accidentally drops a marker on the floor and unsuccessfully reaches for it,” “The adult wants to put magazines into a cabinet, but the doors are closed so that he bumps into it,” “A book slips from a stack as the adult attempts to place it on top of the stack,” and “A spoon drops through a hole and the adult unsuccessfully tries to grasp it through the small hole, ignorant of a flap on the side of the box” (Warneken & Tomasello, 2006). The infants were also presented with control versions of these scenarios where the experimenter did not appear to be having any problems, e.g., purposefully throwing the marker on the floor. In the test condition, the experimenter alternately looked at the object and the infant and verbalized the problem without asking for help. Of the 24 infants tested, 22 helped with at least one task and significantly more infants helped in the test condition compared to the control condition. These results suggest that the infants were able to infer the struggle that the adult was experiencing and were naturally inclined to help. The infants did not receive any praise or rewards for helping and were not asked for help. This suggests that from infancy, mentalizing ability seems to facilitate not just moral evaluations, but an innate moral goodness (Warneken & Tomasello, 2006).

In the studies mentioned, moral evaluations and helping behaviors are the result of being able to infer the intentions of the agent, which would require mentalizing ability. A study by Smetana, Jambon, Conry-Murray, and Sturge-Apple (2012) attempted to empirically test this association between mentalizing ability and moral judgements by observing the relationship between the developments of these two abilities in a longitudinal study over a 1-year period. The results suggested a bidirectional relationship between the development of mentalizing and moral judgement ability, as children who had more mature mentalizing abilities were later able to make more complex moral evaluations, and children who were able to make more complex moral evaluations later developed more mature mentalizing abilities (Smetana et al., 2012).

Failure of Mentalizing Ability

Though social cognition is typically engaged spontaneously when interacting with other human beings, there are certain contexts where this may not take place. Given that mentalizing engages other psychological processes, and is relatively cognitively costly (as captured in the cognitive miser concept, see Fiske & Taylor, 2013), reducing mentalizing may have strategic advantages for the perceiver. As such, the failure to engage mentalizing to a human being may occur for three reasons: (1) as an emotion (empathy) regulation strategy; (2) as a post-hoc justification for immoral behavior; and (3) to facilitate anti-social or atypical social behavior. As such, mentalizing can gate other cognitive processes. For example, in order to be empathetic, a perceiver must engage in mentalizing to allow access into the other person's mind and infer their emotional state. The moral evaluation of others also requires social cognition, since inferences of their intent are central to attribution of moral character. Finally, engaging in behavior typically not reserved for other people, such as a surgeon slicing a person open, or even a nurse administering an injection, may be facilitated by reduced mentalizing of the person since this frees cognitive resources necessary for the task at hand. Therefore, the regulation of mentalizing considerably affects people's behaviors and attitudes towards others.

Given that there are brain regions that reliably engage during social cognition (the SCBN), this provides a brain index for social cognition that can be used to investigate factors that modulate the extent to which people are engaged in social cognition. Here we outline six different social contexts that fail to engage the SCBN and presumably indicate a failure of mentalizing.

Extreme Outgroups

Early studies of dehumanized perception have found evidence of reduced social cognition towards individuals from extreme outgroups. In these studies, extreme outgroups were defined as those that were perceived to be low in both trait warmth and competence, i.e., perceptions of being hostile and of being incapable of acting on those intentions. In a series of fMRI studies, participants were presented with images of individuals that were pre-rated to fall into one of the four quadrants of the SCM (Harris & Fiske, 2006, 2007, 2011). This model posits warmth and competence as the two dimensions of social cognition, with each combination of high or low warmth and competence forming four quadrants. Social groups for each quadrant of the SCM were found to elicit specific emotions in earlier studies, including pride, envy, pity, and disgust (Fiske et al., 2018). For instance, in US samples, Americans elicit pride, business people elicit envy, disabled people elicit pity, and homeless people elicit disgust. These brain imaging studies revealed a consistent pattern of reduced brain activation in the SCBN when participants were presented with individuals from the low warmth, low competence (LW-LC) quadrant, whereas the SCBN was engaged when presented with individuals from the other three quadrants of the SCM. Along with the reduced activation in the SCBN, participants also exhibited greater brain activation in the amygdala and the anterior insula when presented with images from the low-warmth, low-competence quadrant, compared to images from the other three clusters of the SCM. This is in line with the prediction of the SCM for low-warmth, low-competence outgroups to elicit feelings of disgust.

Researchers also found evidence that this reduction of activity in the SCBN in response to the dehumanized extreme outgroups could be moderated. In one such study, participants viewed either images of individuals from extreme outgroups, or individuals from the other three SCM quadrants (Harris & Fiske, 2007). Participants were either asked to make an individuating judgement about the food preferences of the person in the image, or to make a categorical judgement about the age of the person in the image. Again, reduced SCBN activity was observed in response to extreme outgroup members. However, an increase in SCBN activity was observed towards extreme outgroup members in the individuating condition. Additionally, though there was an increase in MPFC activity specifically for both groups, the activation occurred in different regions of the MPFC for extreme outgroup members, compared to members of the other outgroups: The individuating condition increased activation in a dorsal area of the MPFC for extreme outgroup members, while an increase in activation was observed in the ventral area of the mPFC for the other outgroup members. This study provides evidence that while there may be a lack of spontaneous social cognition towards extreme outgroups, a change in context can trigger the re-humanization of extreme outgroup members.

Empathy Avoidance

Another motive for the downregulation of social cognition may be the desire to protect oneself from the negative consequences of empathizing with others who are having negative emotional experiences. For example, empathizing with a homeless person on the street may result in feeling some of their pain and even feelings of guilt if one chooses not to help, or if one feels that their help cannot make a significant difference to the person's life. In order to empathize with someone, it is necessary to mentalize that person to know and understand their psychological experience. The engagement of social cognition therefore acts as a gateway mechanism for empathy. Consequently, the downregulation of social cognition may act as a mechanism to prevent emotional exhaustion in this context by short-circuiting empathy.

A study by Cameron, Harris, and Payne (2016) found evidence supporting this theory. In this study, participants read vignettes of stigmatized versus non-stigmatized target individuals before reporting the level of emotional exhaustion they would anticipate experiencing if they were to help the target. In this study, the researchers measured the level of mentalizing elicited when participants imagined interacting with the target by counting the number of mental state verbs used when participants described a typical day in the life of the target. This study revealed that the effect of stigma on the level of mentalizing was mediated by anticipated emotional exhaustion: Participants anticipated greater emotional exhaustion from the stigmatized targets, resulting in less mentalizing. However, this relationship between stigma and mentalizing held only when participants anticipated high levels of emotional exhaustion. Here the failure to mentalize seems to act as a coping mechanism to protect one's own emotions and resources. However, this is not limited to the context of interactions with stigmatized outgroups. There is also evidence that the downregulation of social cognition can act as a protective mechanism in this way in the medical context.

Medical Context

Evidence from studies investigating the health professionals in the medical context seem to suggest that the downregulation of social cognition, especially the regulation of empathy, can act as a protective mechanism against burnout and emotional fatigue resulting from repeated exposure to the suffering of patients. Gleichgerrcht and Decety (2013) conducted a survey investigating the factors that result in the negative consequences of empathy in qualified physicians in the clinical setting. They measured the relationship between empathy and emotional exhaustion in physicians, including both measures of emotional empathic responses and perspective-taking, which was used as a measure of mentalizing. The more emotional empathic response describes empathy involving emotion recognition and emotional contagion, previously associated with the mirror neuron system (Shamay-Tsoory, 2011; Shamay-Tsoory, Aharon-Peretz, & Perry, 2009). Measures of emotional exhaustion included burnout, secondary traumatic stress, and compassion satisfaction—the positive experience of helping others in distress. The results of the study suggested that physicians who experienced compassion satisfaction and those with compassion fatigue both exhibited high levels of empathic concern. However, physicians with lower stress and burnout scored higher on mentalizing—the cognitive precursor of empathy—and lower on personal distress compared to those who suffered from compassion fatigue. Thus, this study revealed a dissociation between the effects of the more emotional empathic response and the more cognitive mentalizing response. The results further suggest different consequences of either compassion fatigue or compassion satisfaction depending on whether the individual is able to perform the mentalizing required to separate the self from the emotions of the patient.

Experienced physicians with greater compassion satisfaction had greater scores for empathic concern and mentalizing but lower scores for personal distress, suggesting that over time, physicians are able to develop this ability to maintain some emotional distance to patients while simultaneously maintaining empathic concern. In contrast, physicians with compassion fatigue had lower scores of mentalizing ability, suggesting that the distress of the patients may be feeding into their own experience of emotion due to an inability to make the emotional distinction between self and other.

An EEG study by Decety, Yang, and Cheng (2010) also investigated the regulation of empathy, specifically empathy for pain in physicians. Typically, when observing pain in others, brain pathways for pain in the observer are also activated. Decety et al. (2010) found evidence of this in the control sample of their study;

when presented with images of another individual either being pricked by a needle or touched with a cotton bud, controls displayed differential N110 and P300 even related potentials (ERPs) between the pain and no-pain conditions. This was consistent with the participants' subjective ratings of pain intensity between the two conditions. However, physicians did not show any differentiation in N110 and P300 ERPs between the pain and no-pain conditions and reported lower subjective ratings of pain intensity compared to controls. These results suggest that even at the sensory level, there is a downregulation of empathy in physicians. The downregulation of the emotional empathic response may help to free cognitive resources, enabling physicians to perform more effectively, leading to positive emotional outcomes. Thus the downregulation of mentalizing may act as a gateway mechanism facilitating this regulation of emotional empathy.

Labor Market

Another context where the downregulation of mentalizing may be beneficial is in decision-making in the economic context. Harris, Lee, Capestany, and Cohen (2014) conducted an fMRI study investigating the effect of the commodification of people in the context of a labor market. In this study, participants took part in a virtual lab-based labor market where they were asked to purchase players to play on their behalf in a time estimation game. Before the scanning session, images of 60 players were presented to the participants along with their performance ratings, and price based on these ratings. Participants then purchased five players to play on their behalf. In the scanner, participants subsequently revalued both their purchased players and other non-purchased players, with the option of assigning new prices to players. Participants were found to elicit reduced activation of the SCBN when revaluing the players they had purchased. This reduction in SCBN activity also predicted revaluation of purchased players. On the other hand, the revaluation of non-purchased players was predicted by medial orbital frontal cortex (MOFC), which has been implicated in traditional, non-social valuation and willingness-to-pay. This region of the medial frontal cortex is not typically engaged during mentalizing tasks since it sits more ventral to other MPFC regions implicated in mentalizing. These results suggest that the reduction of social cognition may be a functional mechanism involved in making effective evaluative judgements. In economic contexts such as a labor market, it may be beneficial to employers to be flexibly able to set aside social cognition in order to make more rational decisions. For example, when making a hiring decision for a technical position, the temporary downregulation of social cognition may prevent social information from distracting the decision maker from selecting the most skilled applicant. However, this may result in negative outcomes for employees since they are treated as commodities and their psychological well-being is less salient.

Objectification

Another context where we see a reduction in mentalizing is during sexual objectification. Studies have used the body-inversion effect to investigate the objectification of people at a cognitive level. In these studies, the participants view a series of stimuli before being presented with these stimuli a second time alongside new stimuli, and are asked to identify which stimuli they have viewed previously. The inversion effect is the tendency for people to make more recognition errors when presented with an upside-down image compared to an upright image. Usually, an inversion effect is present for human bodies and faces. However, the inversion effect is not usually present when viewing objects (Bernard, Gervais, Allen, Delmée, & Klein, 2015). This is due to the way the two types of images are processed; human bodies and faces are processed configurally—the spatial relationship between features is important for the recognition of the body or face. However, objects are processed analytically—features are processed independently of each other and inversion does not affect recognition (Bernard et al., 2015).

EEG studies have identified a brain correlate of the inversion effect, an amplified N170 ERP (Bernard et al., 2018). As bodies and faces are processed configurally, usually an amplified N170 would be observed in individuals perceiving an inverted image of a human body or face, but no N170 amplification would be present when viewing inverted objects (Bernard et al., 2019). These inversion studies found that an inversion effect is present behaviorally and in the brain when participants were presented with fully clothed human bodies. However, when presented with images of sexualized women, and in one study also sexualized men (Bernard et al., 2019), either with revealing clothing or suggestive poses (Bernard et al., 2019), no inversion effect was produced by participants (Bernard et al., 2015, 2018, 2019; Xiao, Li, Zheng, & Wang, 2019). This suggests that analytical processing is being used for images of sexualized women, similar to object perception.

This elimination of the inversion effect by objectification also seems to be altered by context. Bernard et al. (2019) found no objectification of men and women in revealing clothes but found that men and women in suggestive poses were objectified. Xiao et al. (2019) found that power mediates this effect, finding a positive association between power of the perceiver and the objectification of sexualized women, but not of men. Cikara, Eberhardt, and Fiske (2011) found more direct evidence of the reduced engagement of the SCBN in the brain of the perceiver of objectified targets. In this study, Cikara et al. (2011) found that in men high in hostile sexism scores were associated with reduced activity in the SCBN when perceiving images of sexualized women. Moreover, the researchers also conducted an implicit association test (IAT) to investigate the association between sexualized women/non-sexualized women and first-person action verbs/third-person action verbs. Dehumanized perception tends to be associated with assigning less agency to a social target. Therefore, dehumanized or objectified targets would be associated more with first-person action verbs, as the social target would be more likely to be perceived as the object of an action, whereas social targets that have been assigned agency would be more associated with third-person action verbs (Cikara et al., 2011). Men who scored high on hostile sexism were quicker to associate sexualized women to first-person action verbs, and clothed women to third-person action verbs. This result suggests that highly hostile sexist men attribute less agency, and therefore, less humanity to sexualized women. This may in turn mediate reduced social cognition towards sexualized women, resulting in their objectification.

Social Avoidance

Individual differences also play an important role in mediating the effect of context on social cognition. Beyer, Münte, Erdmann, and Krämer (2014), investigated the role of brain mechanisms underlying social cognition in the emotional reactivity to threat, and how these mechanisms relate to the behavioral and brain effects of provocation on aggression. This study focused on the individual differences in emotional reactivity to threat, which was measured using the fear potentiated startle response (FP). Previous research suggested that those with an increased (FP) tend to score high on harm avoidance and are more likely to feel threatened by harmful stimuli (Beyer, Münte, Erdmann, & Krämer, 2014). Therefore, participants with an increased FP (indicating high emotional reactivity to threat) should be less likely to show aggression. This is because though fear promotes a fight or flight response, the avoidance response is more likely to be used by people who are highly threatened over time and manage to remain within the confines of social convention.

The researchers elicited and measured aggression using the Taylor Aggression Paradigm (TAP). In this task, participants were matched with two confederate opponents on a simple reaction time game. If the participant won, they chose the punishment for the opponent. However, if the participant lost, they would receive punishment determined by their opponent. Punishment was administered by a loud buzzer, and participants chose the volume of this noise blast. In this study, one opponent was provoking, consistently administering high volumes as punishment, while the other was non-provoking, consistently administering low volumes as punishment. Beyer, Münte, Erdmann, and Krämer (2014) found no relationship between FP and aggression. However, they did find a negative correlation between activity in the SCBN and FP in participants selecting punishments for a provoking opponent. Participants with increased FP engaged the SCBN less when interacting with provoking opponents, while participants who exhibited low FP (indicating low emotional reactivity) maintained engagement of the SCBN during their interactions with provoking opponents. These results illustrate another case of reduced social cognition engagement being used as a protective mechanism against negative affect, in this case among individuals especially sensitive to threatening events. Participants with high emotional reactivity to threat may be triggered to suppress mentalizing ability, to protect themselves against the negative affect involved in confronting a threatening opponent.

No direct relationship was found between the regulation of mentalizing ability and aggression in the Beyer, Münte, Erdmann, and Krämer (2014) study; however, evidence of an indirect relationship between the two factors was found in another study by Beyer, Münte, and Krämer (2014). These researchers conducted an fMRI study investigating the relationship between social exclusion and aggression, and the possible mediating role of social cognition between these two factors. In this study, participants played a virtual ball tossing game (cyberball) against two confederate opponents, where the participants were purposefully excluded from the game by the confederate players. Subsequently during scanning, participants completed the Taylor aggression paradigm against the same confederate opponents, and an empathy paradigm where participants viewed neutral or emotional scenes. Beyer, Münte, Erdmann, and Krämer (2014) found that the participants who were excluded showed increased SCBN and mirror neuron system activity, including the bilateral superior gyrus, middle and inferior temporal gyrus, precuneus and right precentral gyrus, when viewing the emotional scenes in the empathy paradigm. Aggression in the TAP was correlated with activation in the inferior temporal gyrus and the right precentral gyrus—regions which showed increased activity in the social exclusion condition, despite there being no direct relationship between social exclusion and aggression in the task (Beyer, Münte, Erdmann, & Krämer, 2014). These results suggest that the disengagement of mentalizing ability acts as a mediating mechanism between social exclusion and aggression.

Virtual Violence

Though studies have not found a direct relationship between aggression and reduced social cognition, studies investigating aggression while playing violent video games have found reduction of activity in the SCBN when engaging in violent gameplay, compared to non-violent gameplay (Mathiak & Weber, 2006; Weber, Ritterfeld, & Mathiak, 2006). While this may be interpreted as reduction of activity in the SCBN acting as a facilitating mechanism for aggression, it may also be acting as a protective mechanism against the negative effects of empathy. Participants in this paradigm were experienced video game players playing a violent first-person shooter game (Call of Duty) against familiar others. Perhaps reduced SCBN, observed right before participants pulled the trigger to kill the avatar of their opponent, was necessary to facilitate such behavior. Thoughts about the mind of the familiar opponent may have delayed such behavior, perhaps providing brain evidence for the classic line in a famous Spaghetti Western *The Good, The Bad, and The Ugly*: “When you come to shoot, shoot, don’t talk.”

Future Directions

There are additional brain regions beyond those identified in the SCBN network that contribute to ancillary processes of social cognition. Of these, cognitive control and executive function are particularly important in social cognition, as social cognition is flexible and regulated in a context-dependent manner. Bock, Gallaway, and Hund (2015) found evidence for an association between cognitive flexibility and theory of mind ability in children over 7 years old, with cognitive flexibility predicting social understanding beyond the effect of age. Rubin, Watson, Duff, and Cohen (2014) also suggested an association between flexible cognition and social cognition via the hippocampus. The hippocampus has been implicated in flexible cognition, in addition to the ACC and DLPFC, as the region responsible for processing relational memories, crucial to both social cognition and executive function. Damage to the hippocampus results in deficits in both executive function and social functions such as empathy and making character judgements (Rubin et al., 2014). This suggests that understanding the role of executive function and cognitive control during flexible social cognition is key to future research on mentalizing failures.

While there have been many studies investigating the mechanisms of cognitive control and its brain correlates, the cognitive control of social cognition has been less well defined. Previous research into cognitive control has identified activity in the ACC and the DLPFC as the brain correlates of cognitive control. The DLPFC has been implicated in top-down cognitive control, while the ACC has been associated with bottom-up processes, monitoring and providing feedback on performance (MacDonald, Cohen, Stenger, & Carter, 2004; McDonald, Cohen, Stenger, & Carter, 2000). Though there have not been many studies of the cognitive control of social cognition specifically, studies of the cognitive control of empathy and the context-dependent regulation of emotions could provide some insight into the cognitive and brain mechanisms that may be involved in the control of social cognition.

Satpute, Badre, and Ochsner (2014) conducted an fMRI study investigating the brain regions involved in the controlled retrieval and selection of social information. They used a semantic judgement task to test these control processes, which are necessary in the process of making social judgements. The researchers found that controlled retrieval of social information engaged the VLPFC, while the selection of goal-relevant social information engaged two regions within the social cognition network—the dorsal MPFC and the TPJ. This suggests that cognitive control may not be a unitary construct since different brain regions mediate the selection of different facets of social information. Further research is necessary to determine whether similar distinctions are observed when attempting to control different social cognitive mechanisms such as mentalizing, perspective-taking, and empathy.

Melloni, Lopez, and Ibanez (2014) proposed the context network model describing the mechanisms and brain regions involved in the context-dependent regulation of empathy. These researchers implicate the MPFC and other frontal regions including the ACC in the integration of contextual cues for the context-dependent regulation of empathy for pain, while temporal regions are implicated in target-specific value judgements based on context. Some of these studies implicate regions of the SCBN in regulatory processes. More research is required to identify whether social cognition self-regulates depending on context, or if the more traditional cognitive control regions are involved in the control of social cognition.

Implications

The effect of context on flexible social cognition may have important implications in policymaking. Policies are created to guide decision-making in order to achieve a desired outcome. Therefore, it would be useful for policymakers to understand the role of flexible social cognition in decision-making. This is especially relevant as the way policies are framed has the potential to change the perceived context of the organization. As we have discussed previously, the perceived context of a situation influences levels of social cognition engagement. This in turn has the potential to influence people's perceptions, attitudes, and behaviors towards each other. This may have the potential to help policymakers consider some of the psychological consequences of policy, which is often overlooked.

One example of an institution where the impact of policy framing on social cognition can be explored is that of the welfare system. Welfare systems tend to be perceived as a social service; however, previous research has found evidence that when welfare is reframed as an unconditional right of citizenship (i.e., a universal basic income system) rather than a conditional payment only to those meeting specific criteria, welfare becomes perceived as an economic good (Calnitsky, 2016). Along with this change in perception, reframing the conditionality of welfare policy resulted in reduced stigma surrounding welfare (Calnitsky, 2016). Based on previous research, we can see that people may engage in social cognition differentially in social versus non-social contexts (Harris et al., 2014). Therefore, framing welfare systems as a financial context versus a social context may have influenced the levels of social cognition engagement, thereby influencing the perceptions, attitudes, and behaviors elicited towards welfare and recipients. Further research is required to investigate the way in which policy framing can influence attitudes and behaviors via social cognition regulation in this and other contexts.

Another context where the regulation of social cognition may have important implications in policy is the medical context. Evidence from previous studies suggests that the downregulation of social cognition (i.e., the failure to mentalize) may act as a protective mechanism against compassion fatigue and burnout (Gleichgerrcht & Decety, 2013). Of course, social cognition is crucial for tasks such as taking history, building trust and rapport with patients, and treating patients fairly and with dignity. However, being able to flexibly engage and disengage in social cognition could potentially improve overall performance. For example, a surgeon may downregulate social cognition engagement during surgery—where empathizing with the patient could distract the surgeon and re-engage in social cognition when debriefing the patient post operation—where it is important for the surgeon to empathize with the patient. Developing a policy that creates a context where physicians can develop and hone their flexible social cognition ability to adapt to various tasks may improve overall performance.

Research on the context-dependent regulation of social cognition could also have important implications in economics. Recent studies of the modulation of social cognition have pointed to flexible social cognition as a mechanism in rational decision-making, especially in economic contexts (Harris et al., 2014). Although in some areas of economics social learning has been associated with irrational behavior such as herding in financial markets (Cipriani & Guarino, 2009), there are also contexts within economics where social cognition is important to make a rational decision leading to a profit maximizing outcome when working with others, for example in labor markets (Harris et al., 2014) or investment behavior (Lee & Harris, 2013). Having an idea where the downregulation of social cognition would be helpful or harmful within the economic context could help organizations to develop economic policies that leverage this ability to flexibly regulate social cognition levels in order to make more profitable decisions with better outcomes.

References

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. *Trends in Cognitive Sciences, 4*(7), 267–278.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews. Neuroscience, 7*, 268. https://doi.org/10.1038/nrn1884

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology, 41*(3), 258.

Bernard, P., Gervais, S. J., Allen, J., Delmée, A., & Klein, O. (2015). From sex objects to human beings: Masking sexual body parts and humanization as moderators to women's objectification. *Psychology of Women Quarterly, 39*, 432. https://doi.org/10.1177/0361684315580125

Bernard, P., Hanoteau, F., Gervais, S., Servais, L., Bertolone, I., Deltenre, P., & Colin, C. (2019). Revealing clothing does not make the object: ERP evidences that cognitive objectification is driven by posture suggestiveness, not by revealing clothing. *Personality and Social Psychology Bulletin, 45*, 16. https://doi.org/10.1177/0146167218775690

Bernard, P., Rizzo, T., Hoonhorst, I., Deliens, G., Gervais, S. J., Eberlen, J., … Klein, O. (2018). The neural correlates of cognitive objectification: An ERP study on the body inversion effect associated with sexualized bodies. *Social Psychological and Personality Science, 9*(5), 550–559.

Beyer, F., Münte, T. F., Erdmann, C., & Krämer, U. M. (2014). Emotional reactivity to threat modulates activity in mentalizing network during aggression. *Social Cognitive and Affective Neuroscience, 9*, 1552. https://doi.org/10.1093/scan/nst146

Beyer, F., Münte, T. F., & Krämer, U. M. (2014). Increased neural reactivity to socio-emotional stimuli links social exclusion and aggression. *Biological Psychology, 96*, 102. https://doi.org/10.1016/j.biopsycho.2013.12.008

Bock, A. M., Gallaway, K. C., & Hund, A. M. (2015). Specifying links between executive functioning and theory of mind during middle childhood: Cognitive flexibility predicts social understanding. *Journal of Cognition and Development, 16*, 509. https://doi.org/10.1080/15248372.2014.888350

Calnitsky, D. (2016). "More Normal than Welfare": The mincome experiment, stigma, and community experience. *Canadian Review of Sociology, 53*, 26. https://doi.org/10.1111/cars.12091

Cameron, C. D., Harris, L. T., & Payne, B. K. (2016). The emotional cost of humanity: Anticipated exhaustion motivates dehumanization of stigmatized targets. *Social Psychological and Personality Science, 7*(2), 105–112. https://doi.org/10.1177/1948550615604453

Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: A review of its functional anatomy and behavioural correlates. *Brain, 129*(3), 564–583.

Cikara, M., Eberhardt, J. L., & Fiske, S. T. (2011). From agents to objects: Sexist attitudes and neural responses to sexualized targets. *Journal of Cognitive Neuroscience, 23*, 540. Retrieved from http://iat.princeton.edu/iat/

Cipriani, M., & Guarino, A. (2009). Herd behavior in financial markets: An experiment with financial market professionals. *Journal of the European Economic Association, 7*(1), 206–233.

D'argembeau, A., Collette, F., Van der Linden, M., Laureys, S., Del Fiore, G., Degueldre, C., … Salmon, E. (2005). Self-referential reflective activity and its relationship with rest: A PET study. *NeuroImage, 25*(2), 616–624.

Decety, J., & Jackson, P. L. (2006). A social-neuroscience perspective on empathy. *Current Directions in Psychological Science, 15*, 54. https://doi.org/10.1111/j.0963-7214.2006.00406.x

Decety, J., Yang, C. Y., & Cheng, Y. (2010). Physicians down-regulate their pain empathy response: An event-related brain potential study. *NeuroImage, 50*(4), 1676–1682. https://doi.org/10.1016/j.neuroimage.2010.01.025

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878–902. https://doi.org/10.1037/0022-3514.82.6.878

Fletcher, P. C., Happe, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S., & Frith, C. D. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition, 57*(2), 109–128.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50*, 531. https://doi.org/10.1016/j.neuron.2006.05.001

Frith, U., & Frith, C. (2001). The biological basis of social interaction. *Current Directions in Psychological Science, 10*(5), 151–155.

Fiske, S. T., & Taylor, S. E. (1991). *McGraw-Hill series in social psychology. Social cognition (2nd ed.)*. Mcgraw-Hill Book Company.

Fiske, S. T., & Taylor, S. E. (2013). *Social cognition: From brains to culture.* Sage.

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences, 7*(2), 77–83.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*, 165. https://doi.org/10.1016/0010-0277(95)00661-H

Gleichgerrcht, E., & Decety, J. (2013). Empathy in clinical practice: How individual dispositions, gender, and experience moderate empathic concern, burnout, and emotional distress in physicians. *PLoS One, 8*(4), 1–12. https://doi.org/10.1371/journal.pone.0061526

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*(7169), 557.

Hamlin, J. K., Wynn, K., Bloom, P., & Mahajan, N. (2011). How infants and toddlers react to antisocial others. *Proceedings of the National Academy of Sciences, 108*(50), 19931–19936.

Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the lowest of the low: Neuroimaging responses to extreme out-groups. *Psychological Science, 17*, 847. https://doi.org/10.1111/j.1467-9280.2006.01793.x

Harris, L. T., & Fiske, S. T. (2007). Social groups that elicit disgust are differentially processed in mPFC. *Social Cognitive and Affective Neuroscience, 2*(1), 45–51. https://doi.org/10.1093/scan/nsl037

Harris, L. T., Lee, V. K., Capestany, B. H., & Cohen, A. O. (2014). Assigning economic value to people results in dehumanization brain response. *Journal of Neuroscience, Psychology, and Economics, 7*(3), 151–163. https://doi.org/10.1037/npe0000020

Harris, L. T., Todorov, A., & Fiske, S. T. (2005). Attributions on the brain: neuro-imaging dispositional inferences, beyond theory of mind. *Neuroimage, 28*(4), 763–769.

Harris, L. T., & Fiske, S. T. (2011). Perceiving humanity or not: A social neuroscience approach to dehumanized perception. In A. Todorov, S. T. Fiske, & D. A. Prentice (Eds.), *Social neuroscience: Toward understanding the underpinnings of the social mind* (pp. 123–134). Oxford University Press.

Iacoboni, M., Lieberman, M. D., Knowlton, B. J., Molnar-Szakacs, I., Moritz, M., Throop, C. J., & Fiske, A. P. (2004). Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. *NeuroImage, 21*(3), 1167–1173.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist, 28*(2), 107–128. https://doi.org/10.1037/h0034225

Lee, V. K., & Harris, L. T. (2013). How social cognition can inform social decision making. *Frontiers in Neuroscience, 7*(7), 1–13. https://doi.org/10.3389/fnins.2013.00259

MacDonald, A., Cohen, J., Stenger, A. V., & Carter, C. (2004). Functional double dissociation of dorsolateral prefrontal cortex and anterior cingulate cortex in cognitive control. *NeuroImage, 11*. https://doi.org/10.1016/s1053-8119(00)90936-4

Mars, R. B., Neubert, F.-X., Noonan, M. P., Sallet, J., Toni, I., & Rushworth, M. F. S. (2012). On the relationship between the "default mode network" and the "social brain". *Frontiers in Human Neuroscience, 6*, 189. https://doi.org/10.3389/fnhum.2012.00189

Mathiak, K., & Weber, R. (2006). Toward brain correlates of natural behavior: FMRI during violent video games. *Human Brain Mapping, 27*, 948. https://doi.org/10.1002/hbm.20234

McDonald, W. A., III, Cohen, D. J., Stenger, A. V., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science, 288*(June), 1835–1838.

Melloni, M., Lopez, V., & Ibanez, A. (2014). Empathy and contextual social cognition. *Cognitive, Affective, & Behavioral Neuroscience, 14*, 407. https://doi.org/10.3758/s13415-013-0205-3

Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The enigmatic temporal pole: A review of findings on social and emotional processing. *Brain, 130*(7), 1718–1731.

Rubin, R. D., Watson, P. D., Duff, M. C., & Cohen, N. J. (2014). The role of the hippocampus in flexible cognition and social behavior. *Frontiers in Human Neuroscience, 8*, 742. https://doi.org/10.3389/fnhum.2014.00742

Ruby, P., & Decety, J. (2004). How would you feel versus how do you think she would feel? A neuroimaging study of perspective-taking with social emotions. *Journal of Cognitive Neuroscience, 16*(6), 988–999.

Satpute, A. B., Badre, D., & Ochsner, K. N. (2014). Distinct regions of prefrontal cortex are associated with the controlled retrieval and selection of social information. *Cerebral Cortex, 24*(5), 1269–1277. https://doi.org/10.1093/cercor/bhs408

Schilbach, L., Eickhoff, S. B., Rotarska-Jagiela, A., Fink, G. R., & Vogeley, K. (2008). Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the "default system" of the brain. *Consciousness and Cognition, 17*(2), 457–467.

Schiller, D., Freeman, J. B., Mitchell, J. P., Uleman, J. S., & Phelps, E. A. (2009). A neural mechanism of first impressions. *Nature Neuroscience, 12*(4), 508.

Shamay-Tsoory, S. G. (2011). The neural bases for empathy. *The Neuroscientist, 17*(1), 18–24.

Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: A double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain, 132*(3), 617–627.

Smetana, J. G., Jambon, M., Conry-Murray, C., & Sturge-Apple, M. L. (2012). Reciprocal associations between young children's developing moral judgments and theory of mind. *Developmental Psychology, 48*(4), 1144.

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D., & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in cognitive sciences, 15*(7), 319–326.

Throop, C. J., Iacoboni, M., Lieberman, M. D., Moritz, M., Knowlton, B. J., Fiske, A. P., & Molnar-Szakacs, I. (2004). Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. *NeuroImage, 21*(3), 1167–1173. https://doi.org/10.1016/j.neuroimage.2003.11.013

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30*(3), 829–858.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science, 311*(5765), 1301–1303.

Weber, R., Ritterfeld, U., & Mathiak, K. (2006). Does playing violent video games induce aggression? Empirical evidence of a functional magnetic resonance imaging study. *Media Psychology, 8*, 39. https://doi.org/10.1207/S1532785XMEP0801_4

Wiggs, C. L., Weisberg, J., & Martin, A. (1998). Neural correlates of semantic and episodic memory retrieval. *Neuropsychologia, 37*(1), 103–118.

Wojciszke, B. (1994). Multiple meanings of behavior: Construing actions in terms of competence or morality. *Journal of Personality and Social Psychology, 67*(2), 222.

Xiao, L., Li, B., Zheng, L., & Wang, F. (2019). The relationship between social power and sexual objectification: Behavioral and ERP data. *Frontiers in Psychology, 10*, 57. https://doi.org/10.3389/fpsyg.2019.00057

Part III
Theoretical Approaches

Linking Models of Theory of Mind and Measures of Human Brain Activity

Sean Dae Houlihan, Joshua B. Tenenbaum, and Rebecca Saxe

Introduction

Human "Theory of Mind" includes the abilities to recognize, infer, reason about, respond to, predict, cause and avoid causing specific beliefs, desires and emotions in other people. The central questions for cognitive neuroscience about these abilities, are: (1) How do people compute these inferences online? That is, how do people combine current evidence with structured priors in specific situations to explain what others are thinking, predict what they will do next, or choose how to respond? (2) How do people learn the structured priors? That is, what combination of evidence, experience, and innate biases drive the acquisition of the framework theory of other minds that people bring to specific interactions? And (3) How are inference in, and development of, Theory of Mind implemented in the human brain? In this chapter, we consider how existing evidence from human neuroimaging experiments helps to constrain answers to these questions.

Understanding the implementation of Theory of Mind in the brain poses some daunting challenges. Mature human Theory of Mind is likely to be at least partially unique to humans. The human behavioral repertoire of flexible cooperation (including pedagogy) and strategic competition imply that humans have a distinctive kind of social intelligence compared to even our closest primate relatives. To the degree that human Theory of Mind is a function, selectively, of human brains, it raises a methodological problem: the methods that we have to study computation in the brain are dramatically more limited for human brains than for other model systems. All existing non-invasive neuroimaging technologies have limited spatial

S. D. Houlihan · J. B. Tenenbaum · R. Saxe (✉)
Department of Brain and Cognitive Science; Center for Brains, Minds and Machines;
McGovern Institute for Brain Research, Massachusetts Institute of Technology,
Cambridge, MA, USA
e-mail: saxe@mit.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_11

resolution, temporal resolution, and coverage. Nevertheless, we argue that the harder challenge is not methodological but theoretical.

We will consider some recent attempts to link models of Theory of Mind to measurements of human brains, the advances that these attempts support, the limits of those advances, and some of the possible next steps. Most importantly, we need explicit linking hypotheses, computational models of how dynamics of activity in neural populations could implement inferences in (or learning of) logically and causally structured theories. In this chapter, we will mostly just point to the gaps that future linking hypotheses could potentially fill.

How We Infer Others' Mental States

One step in the right direction is to begin with a description of the problem space. What is Theory of Mind, and what is it for? A Theory of Mind is an inferred latent causal structure in another mind. We use Theory of Mind to predict a person's future actions based on our estimates of their unobserved mental states, such as their beliefs (Gergely, Nádasdy, Csibra, & Bíró, 1995; Gershman, Gerstenberg, Baker, & Cushman, 2016; Jern & Kemp, 2015; Wimmer & Perner, 1983). But Theory of Mind is not only used to anticipate behavior. We explain actions after they occur, changing our understanding of a person's expectations, values, costs, habits, and intelligence (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Evans, Stuhlmüller, & Goodman, 2016; Gershman et al., 2016; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Jern & Kemp, 2015; Jern, Lucas, & Kemp, 2017; Kliemann & Adolphs, 2018; Kryven, Ullman, Cowan, & CogSci, 2016). These explanations are themselves value-laden: we use Theory of Mind to make moral judgments of a person's actions and character (Cushman, Sheketoff, Wharton, & Carey, 2013; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015). We track a person's knowledge (and its sources) so we know what to learn from her (Gweon & Asaba, 2018; Mills, 2013; Shafto, Eaves, Navarro, & Perfors, 2012). Our causally structured Theory of Mind shapes how we interpret others' expressions (Anzellotti, Houlihan, Liburd, & Saxe, 2019; de Melo, Carnevale, Read, & Gratch, 2014; Ong, Zaki, & Goodman, 2015) and what antecedents evoked them (Wu, Baker, Tenenbaum, & Schulz, 2018). We use our intuitive theory of other people's minds to design interventions: to plan how best to teach in order to change others' beliefs (Bridgers, Jara-Ettinger, & Gweon, 2019; Gweon, Shafto, & Schulz, 2018), or how best to persuade in order to change their desires.

The best known measure of Theory of Mind abilities is the false belief task (Schaafsma, Pfaff, Spunt, & Adolphs, 2015; Wimmer & Perner, 1983). In a traditional false belief task, the participant observes a character who forms a belief based on direct perceptual access (e.g., "the ball is in the box"); while the character is no longer present, the reality is altered (e.g., the ball is moved to the basket); and then the observer is asked about the character's beliefs ("Where does she think the ball is?") or actions ("Where will she first look for the box?"). False belief tasks thus provide a measure of the observer's ability to separately represent where the ball really is, and where the character thinks the ball is.

False belief tasks are useful, but narrow, measures of Theory of Mind; our intuitive causal Theory of Mind supports richer and more generative inferences that include intentions, desires, knowledge, costs, habits, traits, and emotions. These inferences are not binary, but continuous and probabilistic, and allow for quantitative variability in performance. Consider, for example, a different task. The participant observes a character (the hungry graduate student Holly) moving around an environment with obstacles (walls) to get reward (lunch) from one of three food trucks (Korean, Lebanese, and Mexican). There are two parking spots, so at most two trucks are present on any given day. When Holly leaves her office on this day (point Ⓐ in Fig. 1a), she can see that the Korean truck is parked in the close southwest space. The Lebanese truck is parked in the far spot in the northeast corner, but she does not know that because the wall is blocking her line of sight. Suppose that she walks past the Korean truck and around to the other side of the wall, where she can now see the Lebanese truck (point Ⓑ). She then turns around and goes back to the Korean truck (point ©).

To understand Holly's movements, we rely on the central concept of a plan. If her actions are an approximately rational way to achieve her desires given her expectations and costs, then her actions provide a lot of information about those desires. For example, from the observation that she walked past Korean, saw Lebanese, but selected Korean anyway, observers can infer that Holly prefers Mexican food overall and likes Korean second-best. This is a pretty remarkable inference since

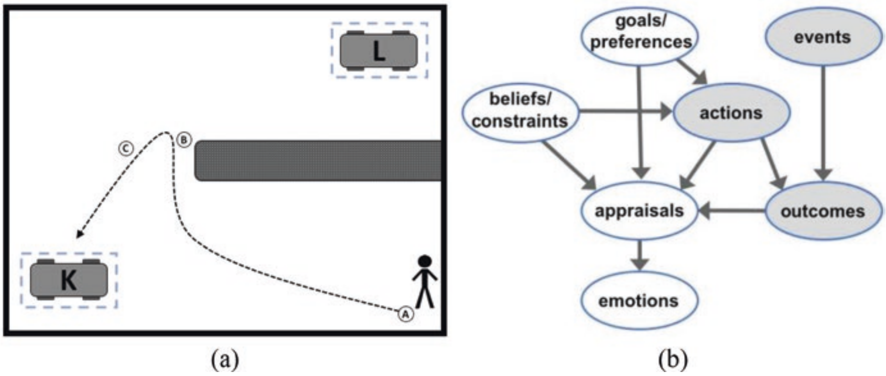


Fig. 1 (**a**) Holly gets lunch. From her initial vantage point Ⓐ, Holly can see the Korean food truck (K) in the southwest parking space, but it is not until she reaches point Ⓑ that she can see past the occluding wall to the second parking space. Observers make graded probabilistic attributions of Holly's beliefs, preferences, costs, rewards, prediction errors, counterfactuals, and emotions at every point along her path. (**b**) Bayesian Theory of Mind, depicted here as a directed acyclic graph. Shaded nodes indicate potentially observable variables, open nodes indicate latent variables, and arrows indicate the causal relationship between variables. As this is a model of people's lay theory of other minds, the model's structure, including implied causal relationships, depicts a hypothesis about people's intuitive reasoning, not a scientific hypothesis about the world itself

observers are systematically inferring Holly's preference for an object that is not present, and that therefore they never observed her choose or even approach. Leveraging this inferred preference, observers can predict Holly's path the next day when the Mexican food truck is parked in the convenient southwestern spot.

In addition to desires, observers can make inferences about Holly's expectations. Because she walked all the way around the building, Holly must have thought it was reasonably likely that the Mexican truck was parked in the northeastern spot. Throughout her path, observers continuously update probabilistic representations of Holly's beliefs and expectations. Inferring Holly's desires and expectations also supports another kind of inference. At the moment she turns the corner and sees the Lebanese truck in the northeastern corner, how does Holly feel? Observers reliably say she feels disappointed: the outcome of her action is going to be less good than she expected (Saxe & Houlihan, 2017).

We can formalize this range of inferences using a probabilistic generative model of Theory of Mind (Fig. 1b). Observers can estimate Holly's desires, recognize the moment her beliefs change, explain past and predict future actions, and anticipate her emotional reactions. Observers' inferences about Holly are well described by the Bayesian Theory of Mind (BToM) model, which supports inferences about rich latent features by probabilistically inverting a generative model of approximately rational agents perceiving, planning, and acting in a dynamic world (Baker et al., 2017; Baker, Saxe, & Tenenbaum, 2009). To make "inverse inferences" (inferences of latent mental contents based on the observation causally connect behaviors) of an agent's beliefs and desires by observing its actions, the observer must have priors over the agent's possible beliefs and desires. To understand Holly's search for lunch, we used a flat prior over beliefs (initially agents think all possible world states are equally probable) and a prior about the structure of desires (each agent has a rank-ordered preference for the three kinds of food). Starting with these priors, BToM jointly infers an agent's beliefs and desires, conditioned on observing the world state and the agent's actions evolving over time. BToM's inferences match human inferences from these scenarios remarkably well, both quantitatively and qualitatively (Baker et al., 2017). Thus BToM offers a quantitative model of how human observers infer a person's specific beliefs and desires, during the temporal evolution of an event, from observations of the world state and the person's actions.

Since the BToM framework was introduced, a classic line of questions about its interpretation has to do with the nature and origins of the generative model relating beliefs and desires to actions for self versus other. Some theorists who favor a "simulation"-like account of action understanding (e.g., other chapters in this volume) have suggested that BToM provides a computational model of this view, if the generative model of action is taken to be the observer's own action planning mechanism. For independent reasons (Saxe, 2005), we favor a "Theory theory"-like account, where Theory of Mind rests on an intuitive theory or mental model of how agents plan. The generative model is an abstract, compressed representation of the causal structure of minds, likely to be simplified, incomplete, or wrong in various ways, but also applicable in situations that the observer might not themselves have any experience planning in or even be able to plan in. This model could of course be applied to predicting or interpreting one's own actions: people represent their own planning during explicit, conscious intuitive reasoning about one's actions, as in rationalization (Cushman, 2019), and people also have an implicit, unconscious "forward model" of their own planning (McNamee & Wolpert, 2019). An abstract schema of how agents plan in general may contain specific sub-models for one's own planning mechanism as well as the plans of specific well-known individuals. For BToM purposes, probabilities of action sequences in a generative model could be evaluated by various means including but not limited to "simulation-based computations" in the technical engineering sense (e.g., Monte Carlo methods for approximate Bayesian inference). But crucially, none of these possibilities reflect the "simulation" accounts of action understanding that some cognitive theorists have proposed, in that the BToM generative model is not implemented in the observer's own planning mechanisms. Only such an interpretation seems to us consistent with the range of inferences—both successful and unsuccessful—that people can carry out with their intuitive Theory of Mind and that we as scientists can model formally and quantitatively using the BToM framework. Nevertheless, the precise relations between BToM computations applied to one's own versus others' actions and thoughts remains an open question.

A more recent challenge for BToM models is to expand the framework (generally called "inverse planning" in reference to the inversion of a forward planning model) to more complex and realistic action plans and environments. In the food trucks examples (and related research on lotteries; Ong et al., 2015), a single actor pursues private goals given individual expectations about the physical world. By contrast, Theory of Mind must also apply to understanding actions in pursuit of social goals (including both direct outcomes for others, Kleiman-Weiner, Ho, Austerweil, Littman, & Tenenbaum, 2016; Kleiman-Weiner, Saxe, & Tenenbaum, 2017; Ullman et al., 2009, and the reputation consequences of actions, Kleiman-Weiner, Shaw, & Tenenbaum, 2017), given expectations that include other people's intentions and actions (Baker, Goodman, & Tenenbaum, 2008; Jern & Kemp, 2014; Kleiman-Weiner et al., 2016; Shum, Kleiman-Weiner, Littman, & Tenenbaum, 2019). Incorporating social interactions will also be necessary to capture a wider array of emotion attributions, such as understanding when a character will feel *pride*, *embarrassment* or *envy* (for example Saxe & Houlihan, 2017). Thus, expanding to more naturalistic settings will necessitate learning an appropriate latent space as well as transformations and computations over that space. Behavioral work has pointed to useful primitive functions (e.g., utilities, reward prediction errors, counterfactuals), but the space of possibilities is large. Discovering the representational abstractions made by neural systems involved in Theory of Mind could heavily constrain the hypothesis space and guide complimentary behavioral modeling approaches. One promising approach is probabilistic program induction, where a hierarchical model learns an inductive bias over inverse planning models like BToM (Lake, Salakhutdinov, & Tenenbaum, 2015; Ong, Soh, Zaki, & Goodman, 2019).

Similarly, a computational model of Theory of Mind should not only match human behavior, but should also suggest hypotheses for neural implementation. We must test how populations of neuronal activity patterns encode the causal structure of another person's inferred expectations, desires and plans. As of now, we still lack any explicit linking hypotheses that could fill this gap. But the results emerging from contemporary neuroimaging experiments suggest we are headed in a useful direction.

Neural Basis of Theory of Mind Inferences

When people are thinking about thinking, a group of brain regions is robustly and reliably recruited (Fig. 2), including bilateral temporal parietal junction (RTPJ, LTPJ), precuneus (PC), and medial prefrontal cortex (MPFC) (Saxe & Powell, 2006; for reviews see Koster-Hale & Saxe, 2013; Saxe & Young, 2013; Schurz, Radua, Aichhorn, Richlan, & Perner, 2014; Spunt, Kemmerer, & Adolphs, 2015). These brain regions, sometimes called the "Theory of Mind network" show high hemodynamic responses to evocations of characters' mental states, compared to evocations of physical states of the world, in non-linguistic cartoons (Gallagher et al., 2000; Sommer et al., 2007) and movies (Jacoby et al., 2016), and in stories presented in writing (Aichhorn et al., 2009; Chan & Lavallee, 2015; Dodell-Feder et al., 2011; Feng, Ye, Mao, & Yue, 2014; Fletcher et al., 1995; Mano, Harada, Sugiura, Saito, & Sadato, 2009; Perner, Aichhorn, Kronbichler, Staffen, & Ladurner, 2006; Saxe & Kanwisher, 2003; Spotorno, Koun, Prado, Van Der Henst, & Noveck, 2012; Vogeley et al., 2001) or aurally (Bedny, Pascual-Leone, & Saxe, 2009; Hervé, Razafimandimby, Jobard, & Tzourio-Mazoyer, 2013; van Ackeren, Casasanto, Bekkering, Hagoort, & Rueschemeyer, 2012), in English (Bedny et al., 2009; Dodell-Feder et al., 2011; Fletcher et al., 1995; Saxe & Kanwisher, 2003), German (Aichhorn et al., 2009; Perner et al., 2006; Vogeley et al., 2001), Dutch (van Ackeren et al., 2012), French (Hervé et al., 2013; Spotorno et al., 2012), Chinese (Chan & Lavallee, 2015; Feng et al., 2014), Japanese (Mano et al., 2009), and American Sign Language (ASL, Richardson et al., 2019). The results from ASL are revealing, because the stimulus (a video of a highly engaging and emotive narrator) is highly social in all conditions; nevertheless activity in this so-called ToM network, in ASL speakers, was high only when the content of the story concerned the mental states of characters. These regions also show much larger responses when thinking about another person's mental states (belief, desires and emotions) than about the internal states of her body (pain, hunger, thirst; Bruneau, Dufour, & Saxe, 2013; Bruneau, Pluta, & Saxe, 2012; Saxe & Powell, 2006; Skerry & Saxe, 2015).

Although there is widespread consensus that TPJ, PC, and MPFC are all robustly recruited during mental state inference, the question of whether any of these brain regions constitute a domain-specific mechanism for Theory of Mind has remained controversial. There are many subtle shades to this controversy, not all of which will be addressed here. One simple question, however, is whether activity during mental state inference actually reflects a different, domain-general cognitive process, which is just incidentally evoked by tasks requiring Theory of Mind. Many such cognitive processes have been hypothesized (Buckner & Carroll, 2007; Lindquist, Wager,

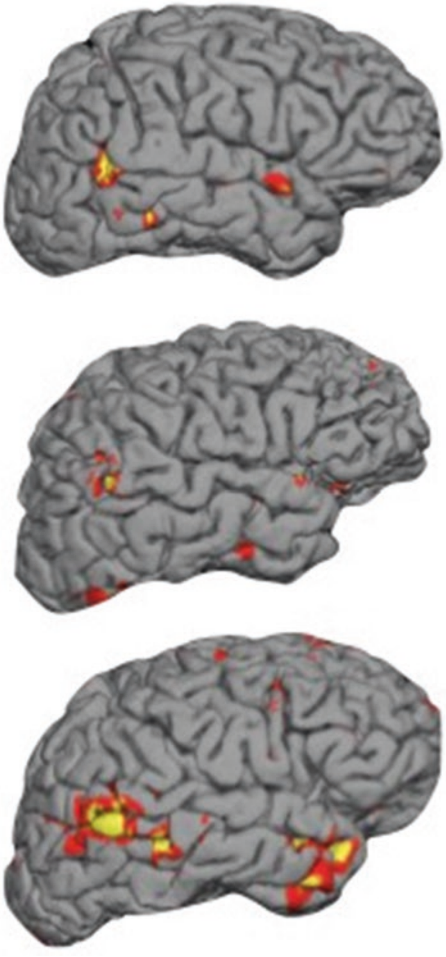


Fig. 2 Thinking about thinking: brain regions commonly recruited in Theory of Mind tasks. (**Left**) Average activation in adults reading stories about others' false beliefs (mental state inference), compared to reading stories about false photographs (non-mental conditions that also requires subjects to represent false or outdated content, e.g., an old photograph that no longer accurately depicts the landscape), overlaid on a template brain (Dodell-Feder, Koster-Hale, Bedny, & Saxe, 2011; Saxe & Kanwisher, 2003). (**Right**) Average activation in adults watching a Pixar animated short film (*Partly Cloudy*), at the moments of salient mental events (e.g., social rejection/isolation, a baby crying and then becoming happy), compared to salient physical events (slapstick physical harm including the protagonist being poked by porcupine quills or bitten by a baby alligator), overlaid on a template brain (Jacoby, Bruneau, Koster-Hale, & Saxe, 2016; Richardson, Lisandrelli, Riobueno-Naylor, & Saxe, 2018). *RTPJ*: right temporo-parietal junction, *PC* precuneus, *vMPFC* ventral medial prefrontal cortex, *dMPFC* dorsal medial prefrontal cortex. Here we collectively term these cortical regions the Theory of Mind network; they are also known as the Mentalizing network

Kober, Bliss-Moreau, & Barrett, 2012; Spreng, Mar, & Kim, 2009). For example, tasks that require reasoning about others' minds might also typically evoke rich episodic memories of one's own similar experiences. Episodic memories do evoke activity in a group of brain regions with a similar distribution across cortex, resembling the so-called "default mode network" (DMN; e.g., Fox et al., 2005; Raichle et al., 2001; Yeo et al., 2011). However, activation of episodic memories cannot explain away the activity in Theory of Mind tasks, because upon closer examination, episodic memory and Theory of Mind recruit activity in almost completely non-overlapping (though spatially nearby and interleaved) cortical regions (DiNicola, Braga, & Buckner, 2019). Standard fMRI methods for data acquisition and analysis blur these neighboring cortical regions together (Braga & Buckner, 2017; Braga, Van Dijk, Polimeni, Eldaief, & Buckner, 2019; Wen, Mitchell, & Duncan, 2019). By collecting much more data within single participants, and then analyzing individual participants separately to preserve idiosyncratic cortical anatomy, DiNicola et al. revealed a striking dissociation between "DMN"s: one involved in memory and projection (future oriented thinking), and the other involved in Theory of Mind.

Other studies have used similar approaches to differentiate the cortical regions involved in Theory of Mind, from nearby regions involved in detecting unexpected events and shifting attention (Scholz, Triantafyllou, Whitfield-Gabrieli, Brown, & Saxe, 2009), perceiving facial and vocal expressions of emotion (Deen, Koldewyn, Kanwisher, & Saxe, 2015), and recognizing social interactions (Isik, Koldewyn, Beeler, & Kanwisher, 2017).

Another function that has been proposed for this cortical network, and especially for TPJ, is narrative comprehension. Responses in TPJ are most robust when a character's mental state is described or evoked in the context of a larger, coherent narrative (Lin et al., 2018). When the narrative coherence is broken, for example by scrambling sentences from a story or scenes from a movie, the response in TPJ is dramatically reduced (Hasson, Yang, Vallines, Heeger, & Rubin, 2008; Lerner, Honey, Silbert, & Hasson, 2011; Lin et al., 2018). An explicit statement of a character's mental state (e.g., "Sarah believes that swimming in the pool is a good way to get cool"), presented in isolation, does not in fact evoke a very strong response in TPJ; narrative context strongly amplifies these regions' response to the same element. An interesting puzzle is therefore how to understand the cognitive and neural dependency between narrative comprehension and Theory of Mind (Jacoby & Fedorenko, 2018; Mar, 2011; Schurz et al., 2014). Is there a cortical network for narrative comprehension, that is typically evoked in ToM tasks but might also be evoked when representing any coherent sequences of events or sentences? Or is there a cortical network for Theory of Mind, which is more robustly recruited when mental states are presented in a coherent narrative context? Although these hypotheses have not been definitively tested, evidence favors the latter interpretation. Coherent expository texts with no mental state content evoke minimal responses in Theory of Mind brain regions (Dodell-Feder et al., 2011; Jacoby et al., 2016); and scrambling these texts has no effect on responses in TPJ (Jacoby & Fedorenko, 2018). Temporally scrambling naturalistic movies (i.e., feature films and TV episodes) does dramatically alter activity in TPJ, but of course these films are designed to evoke rich understanding of characters' minds. Scrambling the order of events plausibly impairs participants' ability to understand and represent the character's more subtle beliefs, desires and emotions.

The Theory of Mind network is thus a set of cortical regions where activity is robustly and selectively evoked by consideration of people's minds. Just finding that a region is selectively active does not address the cognitive or computational questions we posed at the beginning of the chapter. What role do these cortical regions play during online Theory of Mind inferences? One way to investigate is to adapt an approach that has proved highly successful for the ventral visual stream, which is involved in object recognition. A visual image is represented in distinct formats across cortical areas in the ventral visual stream. Low-level stimulus properties like line orientation and shading are linearly decodable from small populations of neurons in early visual areas (e.g., V1) whereas in higher-level regions, the identity of an object becomes linearly decodable and invariant across viewing conditions (DiCarlo, Zoccolan, & Rust, 2012; Kamitani & Tong, 2005; Kourtzi & Kanwisher, 2001; Lafer-Sousa & Conway, 2013; Tanaka, 1993). As information propagates through the ventral pathway, the neural response is reformatted to make features that are relevant to object identity more explicit. Discovering which features of a stimulus can be linearly decoded from each population of neurons can reveal the kinds of representations that those populations support.

By analogy to the visual system, we can ask what features of inferred mental states can be linearly decoded from the patterns of activity in cortical regions. Perhaps amazingly, within Theory of Mind brain regions, different spatial patterns of activity are reliably evoked by descriptions of subtly different mental states, so multi-voxel pattern analyses (MVPA) can be used to find meaningful feature dimensions in the patterns of neural responses to others' mental states. For example, as a first proof of principle, we tested whether patterns of activity in RTPJ differentiate representations of an agent knowingly or unknowingly causing harm. How much a person is blamed for a harmful action (e.g., putting poison in a drink, failing to help someone who is hurt, making an insensitive remark) depends substantially on whether the person reasonably believed that her action would (or would not) cause harm. This aspect of moral evaluation depends disproportionately on the function of RTPJ: causally interfering with activity in the RTPJ shifts moral judgments away from reliance on mental states (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010). Spatial patterns of activity in RTPJ (i.e., which subsets of voxels are relatively more, or relatively less active, within this one region) reliably depend on, and therefore can be used to decode, whether a harmful action was taken with full foreknowledge versus in ignorance. Moreover, individual differences in moral judgment were predicted by individual differences in neural pattern confusability in the RTPJ: people whose RTPJ showed more differentiated patterns of response to intentional versus accidental harms also assigned less blame and greater permissibility to justified accidents (Koster-Hale, Saxe, Dungan, & Young, 2013).

Subsequent research has revealed that the distinction between knowing and unknowing harm is one of many distinctions relevant to Theory of Mind inferences that are decodable from patterns of activity in ToM regions (Koster-Hale et al., 2013, 2017; Koster-Hale, Bedny, & Saxe, 2014; Skerry & Saxe, 2014, 2015; Tamir, Thornton, Contreras, & Mitchell, 2016). The clearest distinction between mental states, based on the patterns of activity evoked in ToM regions, is the valence (or goal-congruence) of the state: did the person get (or expect to get) what she wanted? Although valence is an organizing dimension of all Theory of Mind regions, the representation of this dimension appears to depend disproportionately on MPFC function (Amodio & Frith, 2006; Etkin, Egner, & Kalisch, 2011; Hynes, Baird, & Grafton, 2006; Leopold et al., 2012; Sebastian et al., 2012; Shamay-Tsoory, 2011; Shamay-Tsoory & Aharon-Peretz, 2007; Shamay-Tsoory, Tibi-Elhanany, & Aharon-Peretz, 2006). The population-level activity in MPFC contains abstract, multimodal information about the valence of another person's experience (Chavez & Heatherton, 2014; Chib, Rangel, Shimojo, & O'Doherty, 2009; Chikazoe, Lee, Kriegeskorte, & Anderson, 2014; Kable & Glimcher, 2007; Winecoff et al., 2013). For example, how pleasant the experience is for the protagonist (i.e., the valence of the experience) best explains the pattern of response in MPFC to verbal descriptions of 200 unique emotional events (Skerry & Saxe, 2015). Furthermore, distinct patterns of activity in MPFC are evoked when observing another person (a) make a positive versus negative dynamic facial expression (Harry, Williams, Davis, & Kim, 2013; Peelen, Atkinson, & Vuilleumier, 2010; Said, Moore, Engell, Todorov, & Haxby, 2010; Said, Moore, Norman, Haxby, & Todorov, 2010), (b) make a positive versus negative vocal expression (Peelen et al., 2010), (c) succeed versus fail to complete a goal (like throwing a ball into a net) (Skerry & Saxe, 2014), or (d) get included in versus excluded from a social group (Skerry & Saxe, 2014). This diverse range of stimuli evokes a common multivariate representation of valence such that a linear classifier trained to decode valence based on stimuli from one domain (e.g., stereotypical positive and negative facial expressions) was able to decode valence in a different domain (e.g., animations of expressionless shapes succeeding and failing to accomplish goals) (Skerry & Saxe, 2014).

In addition to distinctions relevant to goals, there are also distinctions relevant to plans or beliefs—including distinctions between planned and unplanned states, and between justified and unjustified beliefs: that is, epistemic features. As in the example of Holly above, observers keep sensitive track of others' expectations, including when and how beliefs change through perception and through inference. RTPJ appears to be differentially important for evaluating other people's beliefs and motivations. The features of another's mind that can be decoded from patterns of activity in RTPJ are epistemic: aspects of the inferred process by which she formed her beliefs. These features include properties of her evidence (e.g., whether her source was something she saw or something she heard; Koster-Hale et al., 2014) and properties of the inference process itself (e.g., whether her conclusions were justified by her evidence or not; Koster-Hale et al., 2017). Evidence justification provides a particularly strong test for features of intuitive epistemology because it is abstract (rather than tied to specific sensory features), context specific (what might be good evidence for one conclusion could be poor evidence for another), and directly related to reasoning about the minds of others (determining whether the agent is a reliable, rational informant; Kovera, Park, & Penrod, 1991; Miene, Borgida, & Park, 1993; Olson, 2003).

As an aside, this distinction between motivation- and valence-biased representations in MPFC, and epistemic representations in RTPJ, may help to resolve a puzzle in the cognitive neuroscience of morality. When a protagonist is described as causing harm knowingly versus unknowingly (e.g., you absolutely knew, versus had no idea about, your cousin's allergy when you served him the peanuts), distinct patterns of activity were observed in RTPJ, and predicted participants' moral judgments of the protagonist (Koster-Hale et al., 2013). By contrast, in a separate experiment, ventral MPFC activity was selective for harmful actions depicted as intentional versus accidental (e.g., deliberately pushing someone versus tripping and falling against them) (Decety, Michalska, & Kinzler, 2012). Furthermore, developmental increase in ventral MPFC selectivity for intentional versus accidental harms was associated with developmental reduction in blame for the accidents (Decety et al., 2012). These two sets of results are compatible when viewed in light of the proposed representational architecture for Theory of Mind: RTPJ contains information about what the protagonist knew or should have known, before acting intentionally (i.e., an epistemic feature), whereas the MPFC is sensitive to whether the action was consistent with the protagonist's goals (i.e., a motivational feature).

There are also other distinctions that can be decoded from patterns of activity, for example separating highly social, high arousal states like playfulness, lust, dominance, and embarrassment, from solitary, low-arousal states like exhaustion, laziness, self-pity and relaxation (Tamir et al., 2016). The distinction here may reflect the mental states of others to which we give resource priority—the ones that inspire our urgent attention—because they drive others' actions and demand our own responses. Interestingly, patterns of brain activity in Theory of Mind regions distinguish between justified and unjustified, but not between true and false beliefs (Koster-Hale et al., 2013). These null results are consistent with the argument above: Theory of Mind concerns the process of making rational inferences from perception and knowledge, not whether the beliefs are true or false. Thus, the distinction between true and false beliefs is not given high priority in the neural representations of Theory of Mind. However, null results in MVPA must always be interpreted with caution. Each fMRI voxel potentially contains hundreds of thousands of neurons so many distinct neural populations are intermingled and indistinguishable at this resolution (Freeman, Brouwer, Heeger, & Merriam, 2011; Op de Beeck, 2010).

There are two general lessons of these studies. First, there is remarkable convergence between the cortical locations of peak selective (univariate) responses and peak (multivariate) information, for representations of others' thoughts. The same cortical areas that show the most selective responses to thinking about mental states (i.e., distinguishing mental state information from other conceptual context, between-domains) also contain the most information about mental states (i.e., distinguishing between one type or feature of mental states and another, within-domains). This convergence between evidence of selectivity and evidence of information content strongly suggests that thinking about thought is implemented in domain-specific representational spaces, distinct from other aspects of conceptual and linguistic processing.

Second, and more importantly, pattern analyses have revealed some of the internal structure of mental state representation. These are the observations that should eventually allow us to test predictions of alternative computational models of mental state inference. Mental states are not simply represented as different from other kinds of states (of the physical world, of the body), there is also an internal structure of similarity, according to which some inferred mental states elicit more similar patterns of activity in ToM brain regions, and others elicit more distinct patterns. The principal dimensions of this internal structure suggest key divisions of labor within mental state inference.

In summary, fMRI evidence suggests an overall organization of representations of mental states. Others' mental experiences are represented as distinct from their bodily experiences; within concepts of other minds, at least two distinct dimensions are made explicit: one separating positive (goal-congruent) from negative (goal-incongruent) states, and at least one other that may track the source and justification of beliefs.

Interpreting Computational Models in Light of Neural Activity

What do these neuroimaging results reveal about the computations underlying Theory of Mind inferences? One proposal, Tamir and Thornton (2018), is that the similarity structure of brain responses directly reveals the substrate of inferences about minds. Using principal components analysis, they find three main organizing dimensions of activity while participants consider the meaning of 60 different terms for states of mind, ranging from "anticipation" and "awe" to "drunkenness" and "disarray" (examples given in Table 1). Tamir and Thornton (2018) argue that representing other minds in this very low dimensional space explains how people are able to make a key type of inference: prediction. Human observers predict that other people's states of mind are more likely to transition between states that are nearby in this abstract 3D space. For example, we expect that a friend now feeling "anxious" will be more likely to feel "sluggish" than "energetic" later (Thornton & Tamir, 2017). Thus, the predicted dynamics of other minds could be captured by trajectories in a low-dimensional neural representation of types of mental states. This idea is exciting because it is a rare attempt to capture the range and richness of mental state inferences, and because of the explicit linking hypothesis between a neural population code and a cognitive inference mechanism.

We suggest an alternative: that the dynamics of mental states must be understood in terms of causally and logically structured relations between mental contents, not simply transition probabilities. Mental state attributions are not likely to be well-described as simply a list of features; rather, they require representations with

Table 1 Example stimuli

Tamir et al.	Skerry and Saxe
Planning:	**Mental:**
"carrying emergency cash" "executing a science experiment" "looking at the weekend's weather" "researching an item to purchase it"	"Lucy and her teammates trained hard in preparation for the upcoming soccer play offs. Their coach told them they had a chance of winning the championship. On the first day of the playoffs, a few fluke plays put Lucy's team down 2 to 0. They lost the game, knocking them out of the playoffs in the first round."
Belief: "listening to a religious service" "confident about an attitude" "reading the Bible" "wearing a lucky charm"	
Opinion: "thinking California is the best state" "personal belief" "finding brunettes more attractive" "recommending a type of music"	"Jordan swore to her roommates that she would keep her new diet. Later, she was in the kitchen getting a glass of water, and took a bite of a cake she had bought for their dinner party the following evening. Jordan's roommates arrived home to find that she had eaten half the cake and broken her diet."
Thought: "putting ideas together" "putting ideas together" "remembering to bring an umbrella" "deciding what to do today" "forming an opinion"	"Jake always avoided the doctor's office because he really disliked needles. One summer, Jake was traveling to Kenya for a project, and was told he needed a series of tests and vaccinations before he could go. He reluctantly called the travel clinic and scheduled an appointment for the following week."
Anticipation: "on the line to ride a rollercoaster" "waiting for a band to go onstage"	
Lust: "feeling horny" "preferring physical to emotional"	**Physical:** "Roger was walking to school when he heard a friend behind him call his name. Roger turned to respond, but just then tripped and stumbled over some wood on the ground. Roger fell forward and impaled his hand on a rusty nail in the wood."
Drunkenness: "drinking alone" "spending time with an alcoholic"	

To capture mental state inferences, Tamir et al. (2016) presented a pair of scenarios and asked participants which would better evoke the associated mental state in another person. For instance, participants indicated whether the mental state "thought" was better evoked by "forming an opinion" or "deciding what to do today". Both scenarios are intended to evoke the associated mental state so there is not a "correct" answer. Skerry and Saxe (2015) showed participants narratives that prompt mentalistic inferences about plans, beliefs, expectations, desires, reactions and emotions, and narratives that prompt inferences of bodily sensations (these Physical stories do not evoke activity in Theory of Mind regions). Using similar analysis techniques, Tamir et al. concluded that a low-dimensional representational space (four dimensions) could sufficiently capture behavioral judgments and neural activity during ToM, while Skerry and Saxe concluded that mental state representations are much higher dimensional (>10 dimensions). One possibility is that Skerry and Saxe's inclusion of richer context, and more specific content, evokes more differentiated cognitive and neural representations of mental states internal structure (Baker et al., 2017, 2009; Davidson, 1963), understood in terms of their computational role within a coherent explanatory theory (cf. theory; Carey, 2009; Gopnik & Wellman, 1994). Any representational similarity analysis operationalizes these representations as a "bag of features," more similar to the way concepts have been defined in prototype theory (i.e., graded categorization based on feature similarity to some category prototype or centroid; Rosch, 1973). This approach contrasts with traditional "mental states," which are composed of an attitude (or evaluative perception) toward a proposition (or content). We cannot ask how a person's belief will influence her next action without knowing: her belief about what? Even a simple propositional attitude (e.g., "The father fears his son will fall out of the tree") is composed of an agent (the father), an attitude (fears), and a propositional content (child falling out of tree), and is causally connected to many other specific mental states (e.g., perceptual evidence of wobbly branches, desires to intervene, conflicting desires to promote independence, and so on). The current vector space models do not encode logical or causal structure (context), and lack compositionally (content). The difference between feeling "playful" versus "serious" might be measurable as the distance between two vectors along one continuous dimension, but the difference between "wanting the ball" versus "wanting to go to the ball," or "wanting to play" and "wanting to go to the play," are different in kind. Different formal structures will likely be required (Baker et al., 2017; Skerry & Saxe, 2015). Relatedly, inferences about beliefs necessarily depend on a rich body of world knowledge (e.g., about trees, and about children), so neural populations specific to Theory of Mind must interface with general-purpose semantic systems. A list of features made explicit by each neural population is not enough to test alternative theories of inference in Theory of Mind.

Consistent with this theoretical perspective, there are already empirical hints that representations in the Theory of Mind network are not low-dimensional. We found that patterns of response in the ToM network, including RTPJ and MPFC, can be used to classify verbal narratives (examples given in Table 1) into 20 distinct emotion labels (e.g., *furious*, *jealous*, *grateful*, *proud*; Skerry & Saxe, 2015). The features that explained significant variance in the neural response are natural components of planning and belief updating, and not all easily captured by the three-dimensional solution: for example, whether the event would be repeated in the future, affected the protagonist's life in the long run, and/or was caused by the protagonist or by other people. We found that a minimum of ten feature dimensions were required to explain the reliable variance in that dataset, and that is still likely to be a substantial underestimate. Just within the representation of "rationality" or the reasons for others' beliefs, we have already discovered more than one dimension. In RTPJ, within a single task and set of stimuli, patterns of activity in RTPJ can be used to decode whether the person's beliefs were formed based on sufficient or insufficient evidence, and whether they were based on visual or auditory evidence—the patterns of activity that distinguished beliefs based on modality versus justification were orthogonal (Koster-Hale et al., 2017). Furthermore, evoking rich and specific mental states requires relatively long and complex stimuli. For example,

> *Ginny's classmate wants to borrow a bike to go mountain biking. Ginny's sister left her bike in the garage when she went off to college. The bike had been in and out of the shop for brake trouble. Ginny believes the bike is fully functional now, since the last time she talked to her sister, the brakes were working fine. Ginny lends her classmate the bike, which turns out not to be fully fixed. Her classmate crashes into a tree due to the defective brakes and loses her two front teeth.*

implies a justified belief and induces a distinct pattern of activity in RTPJ from the pattern induced by replacing the emphasized text with "though the last time she talked to her sister, the brakes were still giving her trouble." By classifying average neural responses to a whole sentence, presented in the context of a longer narrative, we combined many cognitive processes. As a result, classification results must be interpreted as a lower bound on the information available in each region (Kriegeskorte & Kievit, 2013).

In sum, we propose that neural populations within the Theory of Mind network support inference by implementing something like the BToM computations: building and operating over a probabilistic causal model of others' motives, expectations and plans. This proposal remains mostly a promissory note. It is missing specific linking hypotheses for how stimuli (i.e., observed events, verbal narratives) are transformed into neural representations, and how priors are represented and combined with representations of the input (which requires a theory of how neurons encode prior knowledge). To make progress in this research program, it will be necessary to construct at least one, but ideally competing, models of how Theory of Mind inference could work in principle, along with more robust linking hypotheses concerning the neural implementation, and the resulting features that might be detectable at the resolution of fMRI. For many reasons, this may fail. But given the current trajectory of progress, it seems worth a shot.

Neural Basis of Theory of Mind Development

A fundamental component of any hypothesis about Theory of Mind inference must be a representation of structured prior knowledge. Holly's movements around her campus can only reveal her preferences and beliefs in virtue of prior knowledge about human planning—that people typically have a rank-ordered preference for foods, that longer paths are more costly, that beliefs can be updated via direct visual access, and so on. How are these priors acquired, and implemented neurally? Using what we know about the mature ToM network, we can operationalize one part of this question by asking how children come to have cortical regions, in RTPJ, MPFC and elsewhere, that are selectively recruited by reasoning about other minds. Is the dramatic and stereotyped development of Theory of Mind abilities during early childhood associated with functional changes in these regions? Are the functions of these brain regions learned? Are they constrained by biological predispositions, and if so, how?

Classic theoretical debates about social cognitive development have considered two opposing possibilities, arguing that ToM is either instantiated in a distinct domain-specific biological mechanism or is constructed through conversational interactions and social relationships (Carlson & Moses, 2001; Hughes & Devine, 2015; Scholl & Leslie, 2001). By contrast, we suggest that ToM is both; Theory of Mind is acquired by a domain-specific biological mechanism, whose mature function and selectivity is constructed in part through linguistically-mediated transmission of culturally-specific concepts.

As described in the previous section, adults have a highly reliable set of cortical regions that are recruited selectively when reasoning about other minds. Activity in these regions is high when thinking about their thoughts or emotions, but not when considering other features of the same characters, including their physical actions and bodily sensations. We argued earlier that these regions constitute a domain-specific biological mechanism with a selective function in Theory of Mind. The functions of these regions are distinct from other aspects of social cognition very early in development.

In 3 year old children, before they can pass false belief tasks, the ToM regions are functionally correlated with each other and respond to evocations of characters' mental states (Richardson et al., 2018). Activity in RTPJ peaks when characters have a false belief, even in preverbal infants (Hyde, Simon, Ting, & Nikolaeva, 2018). Thus, in some sense ToM regions are predisposed to some function related to Theory of Mind, from very early in development. These early origins are not incompatible with environmental influence. On the contrary, we hypothesize that the specific representations and computations of these regions are shaped during development through conversational interactions and social relationships.

Activity in the RTPJ is particularly selective for thinking about others' thoughts in adults (Bruneau et al., 2012; Dodell-Feder et al., 2011; Jacoby et al., 2016; Lombardo et al., 2010; Mitchell, Banaji, & Macrae, 2005; Saxe & Kanwisher, 2003; Saxe & Powell, 2006; Spunt et al., 2015). Similar to the development of cortical regions specialized for other functions, development of increased selectivity in the RTPJ occurs by the suppression of responses to non-preferred stimuli. For example, selectivity of the fusiform face area (FFA) develops through the suppression of responses to (non-preferred) non-face objects; this suppression is correlated with performance on face recognition tasks (Cantlon, Pinel, Dehaene, & Pelphrey, 2010; Golarai et al., 2007; Gomez et al., 2017). Selectivity of the visual word form area (VWFA) develops through the suppression of responses to (non-preferred) faces (Cantlon et al., 2010), and this suppression predicts literacy and reading expertise (Dehaene et al., 2010; Dehaene-Lambertz, Monzalvo, & Dehaene, 2018). Similarly, selectivity of the RTPJ develops through suppression of responses to other (non-mentalistic) social information (Gweon, Dodell-Feder, Bedny, & Saxe, 2012; Saxe, Whitfield-Gabrieli, Scholz, & Pelphrey, 2009), and correlates with performance on ToM tasks (Gweon et al., 2012). For example, in adults, verbal descriptions of a person's physical appearance, place of origin, or social relationships elicit little activity in RTPJ, compared to descriptions of a person's beliefs, desires and emotions (Gweon et al., 2012; Mitchell et al., 2005; Saxe & Powell, 2006). In young children, all of these different kinds of social cues evoke high responses in RTPJ, compared to non-social controls (e.g., descriptions of the physical environment) (Gweon et al., 2012; Saxe et al., 2009).

In the case of FFA and VWFA, extensive domain-relevant experience precedes the emergence of a selective cortical region. What drives the developmental acquisition of RTPJ selectivity and what role does environmental experience play?. A particularly important source of input that children use to build a Theory of Mind is linguistically rich conversational experience. In conversation, adults use words and sentences to describe their mental states and experiences (Harris, 1992, 2002). Even utterances that do not contain mental state verbs (e.g., "Where is my hat?") frequently provide evidence about another person's beliefs and desires, which then help to interpret behavior (Peterson & Siegal, 2016; Siegal & Peterson, 1994). However, utterances that do include mental state verbs may be a particularly rich source of information: children learn to differentiate mental state concepts (e.g., believe vs. know, want vs. hope, peek vs. stare) from the way adults use these mental state verbs in conversational context (Gleitman, 1990). Indeed, just the existence of these distinct words may be an important source of evidence to children, concerning the structure and kinds of mental state concepts used in their culture.

The clearest evidence that linguistic experience affects ToM development comes from studies of children who are d/Deaf and not exposed natively to a sign language. Many deaf or hard of hearing children are at risk of not learning any language in early childhood because they have limited auditory access to spoken language, and their families do not know sign language at the time of birth (Mitchell & Karchmer, 2004). Deaf children with delayed exposure to sign language show corresponding delays in ToM relative to typically hearing children and deaf children exposed to sign language from infancy (Figueras-Costa & Harris, 2001; Gale, De Villiers, De Villiers, & Pyers, 1996; Peterson & Siegal, 2016; Peterson & Wellman, 2018; Peterson, Wellman, & Liu, 2005; Peterson, Wellman, & Slaughter, 2012; Schick & Hoffmeister, 2001; Schick, de Villiers, de Villiers, & Hoffmeister, 2007; Woolfe, Want, & Siegal, 2002). Hearing parents who learn sign language as a second language exhibit large variability in their use of mental state language, which in turn predicts their deaf children's performance on ToM tasks (Moeller & Schick, 2006).

We therefore tested the effect of delayed access to language on the development of selectivity in RTPJ (Richardson et al., 2019). In native signing children, the RTPJ showed selective responses to stories about mental states in the linguistic ToM task. Like native signers, delayed signing children showed high responses to Mental stories ("Jimmy soon realized the pirate didn't know where the treasure was"), but the response in their RTPJ was also high for non-mentalistic social information—like physical appearances and enduring relationships (Social stories: "Old Mr. McFeegle is a gray wrinkled old farmer"; "Sarah and Lori play together on the soccer team"). The reduced selectivity in RTPJ was similar to the response profile previously observed in young children (Gweon et al., 2012; Saxe et al., 2009). Delayed access to ASL correlated with delayed selectivity of RTPJ for mental state information, despite relatively short delays prior to language exposure, and despite being highly proficient in ASL comprehension (matched to native signers) at the time of testing (Richardson et al., 2019).

Conversational experience is not only necessary for acquisition of mental state concepts, it can also be sufficient. The clearest evidence for the sufficiency of conversational exposure comes from the incredible richness of congenitally blind people's knowledge about sight. If first-person experience is necessary to understand others' experiences, blind people should have only a fragmentary, limited, or metaphorical understanding of seeing. But they do not. On the contrary, through conversation and social interaction with sighted people, blind people acquire a rich intuitive theory of sight. Even young blind children know that other people can see with their eyes, and understand for example that objects can be seen from a distance and are invisible in the dark (Bigelow, 1992; Landau & Gleitman, 1985; Peterson, Peterson, & Webb, 2000). By adulthood, congenitally blind people know the meanings of verbs of sight, including fine-grained distinctions between concepts like peer, gaze, and gawk (Bedny, Koster-Hale, Elli, Yazzolino, & Saxe, 2019; Landau & Gleitman, 1985; Lenci, Baroni, Cazzolli, & Marotta, 2013). Finally, the similarity between blind and sighted people's reasoning about sight is evident not just in behavior but also in neural implementation. Like sighted people, blind people recruit RTPJ selectively when thinking about other people's experiences of seeing, but not their experience of bodily states like hunger or nausea (Bedny et al., 2009), and the pattern of neural activity in the RTPJ of both blind and sighted people can decode the source of the character's belief from auditory versus visual evidence (Koster-Hale et al., 2014).

In summary, we propose that during development, children learn a model of the latent causal structure of other minds. This learning occurs through conversational interactions and social relationships, and thus is attuned to the distinctions and structures of other minds that are relevant in the child's cultural context. On the other hand, learning some kind of model of other minds is in a sense biologically prepared by, and preferentially attached to a reliable cortical mechanism and thus appears in the same highly selective regions across individuals, languages, and groups. What is learned by these cortical regions must be not only a division of the domain of minds from other aspects of social life, but also the structured priors (i.e., the framework theory) about how minds work in general that supports specific inferences about one person's beliefs or desires in one particular context. As above, future work is required to define testable linking hypothesis for how development of domain-specific brain regions constitutes the construction of structured priors for inferences.

Future Directions: Linking Neural Measures to Computational Models

For the next step in a deeper understanding of both inference and development of Theory of Mind, we need well-specified hypotheses for how neural dynamics could implement computations over a mental model of latent causal structure. This is a lofty goal, and not unique to Theory of Mind. Other domains of cognitive neuroscience, including the neural basis of language and of intuitive physics, face a similar challenge. The solution to this challenge is unknown, so here we point in some promising future directions.

The first step is to define a range of Theory of Mind inferences that (1) covers the rich and elaborated structure of the intuitive Theory of Mind, and (2) can be well captured by computational models of inferences. We propose that a good starting point is inferences about others' reactions to unfolding events (Ong et al., 2015; Saxe & Houlihan, 2017). Predicting another person's reactions requires a causal model of their mind, because reactions happen when people's expectations, desires, plans and habits meet a dynamic world. For example, when Holly the graduate student sets out looking for lunch, her plans reveal her expectations (where the food trucks will be) and preferences (which cuisines she prefers). At the moment that she turns the corner and sees her least favorite truck parked in the northeast spot, observers infer that Holly can update her expectations based on her perception (an epistemic change). Because changing her expectations about the trucks changes her expected reward in the situation, observers also recognize that Holly is experiencing negative reward prediction error—that is, disappointment (Ong et al., 2015; Wu et al., 2018).

We propose that BToM can be expanded to match human observers' inference about others' emotions (Saxe & Houlihan, 2017). BToM probabilistic generative models are designed to update posterior estimates of a person's preferences and expectations based on her actions, and then compute the consequences of events in terms of the person's achieved utilities (did she get what she wanted), prediction errors (did she get what she expected), counterfactual utilities (what would she have gotten if she chose a different action), and so on. If, as we suggest, these features are core components of Theory of Mind inferences, then they should also provide a good fit to neural activity during those inferences (Skerry & Saxe, 2015). That is, the features computed by BToM could be used as an encoding model (Mitchell et al., 2008; Naselaris, Kay, Nishimoto, & Gallant, 2011) for fMRI responses: an explicit hypothesis about the features represented explicitly in the Theory of Mind brain regions.

Conclusions

How are inference in, and development of, Theory of Mind, implemented in the human brain? Here we argue that Theory of Mind inferences are implemented, at least partially, in distinct and selective cortical regions. Within these regions, neural activity is generally high and sustained, while people think about thoughts, and distinct patterns of population activity contain information about abstract dimensions or features of the inferred mental states, including valence and rationality. The strong selectivity, and presumably the distinct spatial patterns, in these cortical regions emerge reliably during development. However, adult cortical divisions of labor are not fully innately prespecified, but rather emerge in social and cultural context. As yet, there are no testable (let alone competing) models linking the activity in these cortical regions to adequate inferential processes over causal models that can capture the sensitivity of human Theory of Mind. Development of such models is a critical direction for future research.

Acknowledgments We would like to sincerely thank Hilary Richardson for providing Fig. 2. We are grateful for the immensely useful feedback we received from Daniel Nettle, Stefano Anzellotti, and Michael Gilead on an earlier draft of this chapter. This work was supported by the McGovern Institute and the Center for Brains, Minds and Machines (CBMM; funded by NSF STC award CCF-1231216).

References

Aichhorn, M., Perner, J., Weiss, B., Kronbichler, M., Staffen, W., & Ladurner, G. (2009). Temporo-parietal junction activity in theory-of-mind tasks: Falseness, beliefs, or attention. *Journal of Cognitive Neuroscience, 21*(6), 1179–1192.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*(4), 268–277.

Anzellotti, S., Houlihan, S. D., Liburd, S., & Saxe, R. (2019). Leveraging facial expressions and contextual information to investigate opaque representations of emotions. *Emotion*.

Baker, C. L., Goodman, N. D., & Tenenbaum, J. B. (2008). Theory-based social goal inference. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour, 1*(4), 598.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*(3), 329–349.

Bedny, M., Koster-Hale, J., Elli, G., Yazzolino, L., & Saxe, R. (2019). There's more to "sparkle" than meets the eye: Knowledge of vision and light verbs among congenitally blind and sighted individuals. *Cognition, 189*, 105–115.

Bedny, M., Pascual-Leone, A., & Saxe, R. R. (2009). Growing up blind does not change the neural bases of Theory of Mind. *Proceedings of the National Academy of Sciences of the United States of America, 106*(27), 11312–11317.

Bigelow, A. E. (1992). Blind children's ability to predict what another sees. *Journal of Visual Impairment & Blindness, 86*(4), 181–184.

Braga, R. M., & Buckner, R. L. (2017). Parallel interdigitated distributed networks within the individual estimated by intrinsic functional connectivity. *Neuron, 95*(2), 457–471.e5.

Braga, R. M., Van Dijk, K. R. A., Polimeni, J. R., Eldaief, M. C., & Buckner, R. L. (2019). Parallel distributed networks resolved at high resolution reveal close juxta-position of distinct regions. *Journal of Neurophysiology, 121*(4), 1513–1534.

Bridgers, S., Jara-Ettinger, J., & Gweon, H. (2019). Young children consider the expected utility of others' learning to decide what to teach. *Nature Human Behaviour, 16*, 382.

Bruneau, E., Dufour, N., & Saxe, R. (2013). How we know it hurts: Item analysis of written narratives reveals distinct neural responses to others' physical pain and emotional suffering. *PLoS One, 8*(4), 1–9.

Bruneau, E. G., Pluta, A., & Saxe, R. (2012). Distinct roles of the 'Shared Pain' and 'Theory of Mind' networks in processing others' emotional suffering. *Neuropsychologia, 50*(2), 219–231.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences, 11*(2), 49–57.

Cantlon, J. F., Pinel, P., Dehaene, S., & Pelphrey, K. A. (2010). Cortical representations of symbols, objects, and faces are pruned back during early childhood. *Cerebral Cortex, 21*(1), 191–199.

Carey, S. (2009). *The origin of concepts*. Oxford: Oxford University Press.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development, 72*(4), 1032–1053.

Chan, Y.-C., & Lavallee, J. P. (2015). Temporo-parietal and fronto-parietal lobe contributions to theory of mind and executive control: An fMRI study of verbal jokes. *Frontiers in Psychology, 6*, 1405.

Chavez, R. S., & Heatherton, T. F. (2014). Multimodal frontostriatal connectivity underlies individual differences in self-esteem. *Social Cognitive and Affective Neuroscience, 10*(3), 364–370.

Chib, V. S., Rangel, A., Shimojo, S., & O'Doherty, J. P. (2009). Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *Journal of Neuroscience, 29*(39), 12315–12320.

Chikazoe, J., Lee, D. H., Kriegeskorte, N., & Anderson, A. K. (2014). Population coding of affect across stimuli, modalities and individuals. *Nature Neuroscience, 17*(8), 1114–1122.

Cushman, F. (2019). Rationalization is rational. *The Behavioral and Brain Sciences, 43*, 1–69.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition, 127*(1), 6–21.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy, 60*(23), 685–700.

de Melo, C. M., Carnevale, P. J., Read, S. J., & Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. *Journal of Personality and Social Psychology, 106*(1), 73–88.

Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. *Cerebral Cortex, 22*(1), 209–220.

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. *Cerebral Cortex, 25*(11), 4596–4609.

Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Nunes Filho, G., Jobert, A., … Cohen, L. (2010). How learning to read changes the cortical networks for vision and language. *Science, 330*(6009), 1359–1364.

Dehaene-Lambertz, G., Monzalvo, K., & Dehaene, S. (2018). The emergence of the visual word form: Longitudinal evolution of category-specific ventral visual areas during reading acquisition. *PLoS Biology, 16*(3), e2004103.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron, 73*(3), 415–434.

DiNicola, L. M., Braga, R. M., & Buckner, R. L. (2019). Parallel distributed-works dissociate episodic and social functions within the individual. *bioRxiv*, 733048.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. *NeuroImage, 55*(2), 705–712.

Etkin, A., Egner, T., & Kalisch, R. (2011). Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in Cognitive Sciences, 15*(2), 85–93.

Evans, O., Stuhlmüller, A., & Goodman, N. D. (2016). Learning the preferences of ignorant, inconsistent agents. In *30th AAAI Conference on Artificial Intelligence, AAAI 2016* (pp. 323–329). Oxford: University of Oxford.

Feng, S., Ye, X., Mao, L., & Yue, X. (2014). The activation of theory of mind network differentiates between point-to-self and point-to-other verbal jokes: An fMRI study. *Neuroscience Letters, 564*, 32–36.

Figueras-Costa, B., & Harris, P. (2001). Theory of mind development in deaf children: A nonverbal test of false-belief understanding. *The Journal of Deaf Studies and Deaf Education, 6*(2), 92–102.

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J., & Frith, C. D. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition, 57*(2), 109–128.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America, 102*(27), 9673–9678.

Freeman, J., Brouwer, G. J., Heeger, D. J., & Merriam, E. P. (2011). Orientation decoding depends on maps, not columns. *Journal of Neuroscience, 31*(13), 4792–4804.

Gale, E., De Villiers, P., De Villiers, J., & Pyers, J. (1996). Language and theory of mind in oral deaf children. In *Proceedings of the 20th Annual Boston University Conference on Language Development* (pp. 213–224). Somerville, MA: Cascadilla Press.

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: An fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia, 38*(1), 11–21.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*(2), 165–193.

Gershman, S. J., Gerstenberg, T., Baker, C. L., & Cushman, F. A. (2016). Plans, habits, and theory of mind. *PLoS One, 11*(9), e0162246–e0162224.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition, 1*(1), 3–55.

Golarai, G., Ghahremani, D. G., Whitfield-Gabrieli, S., Reiss, A., Eberhardt, J. L., Gabrieli, J. D., & Grill-Spector, K. (2007). Differential development of high-level visual cortex correlates with category-specific recognition memory. *Nature Neuroscience, 10*(4), 512.

Gomez, J., Barnett, M. A., Natu, V., Mezer, A., Palomero-Gallagher, N., Weiner, K. S., … Grill-Spector, K. (2017). Microstructural proliferation in human cortex is coupled with the development of face processing. *Science, 355*(6320), 68–71.

Gopnik, A., & Wellman, H. M. (1994). *The theory. Mapping the mind: Domain specificity in cognition and culture* (p. 257). Cambridge: Cambridge University Press.

Gweon, H., & Asaba, M. (2018). Order matters: Children's evaluation of underinformative teachers depends on context. *Child Development, 89*(3), e278–e292.

Gweon, H., Dodell-Feder, D., Bedny, M., & Saxe, R. (2012). Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. *Child Development, 83*(6), 1853–1868.

Gweon, H., Shafto, P., & Schulz, L. (2018). Development of children's sensitivity to overinformativeness in learning and teaching. *Developmental Psychology, 54*(11), 2113–2125.

Harris, P. L. (1992). From simulation to folk psychology: The case for development. *Mind & Language, 7*(1–2), 120–144.

Harris, P. L. (2002). What do children learn from testimony? In P. Carruthers, S. Stich, & M. Siegal (Eds.), *The cognitive basis of science* (pp. 316–334). Cambridge: Cambridge University Press.

Harry, B., Williams, M. A., Davis, C., & Kim, J. (2013). Emotional expressions evoke a differential response in the fusiform face area. *Frontiers in Human Neuroscience, 7*, 692.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *The Journal of Neuroscience, 28*(10), 2539–2550.

Hervé, P.-Y., Razafimandimby, A., Jobard, G., & Tzourio-Mazoyer, N. (2013). A shared neural substrate for mentalizing and the affective component of sentence comprehension. *PLoS One, 8*(1), e54400.

Hughes, C., & Devine, R. T. (2015). Individual differences in theory of mind from preschool to adolescence: Achievements and directions. *Child Development Perspectives, 9*(3), 149–153.

Hyde, D. C., Simon, C. E., Ting, F., & Nikolaeva, J. I. (2018). Functional organization of the temporal-parietal junction for theory of mind in preverbal infants: A near-infrared spectroscopy study. *The Journal of Neuroscience, 38*(18), 4264–4274.

Hynes, C. A., Baird, A. A., & Grafton, S. T. (2006). Differential role of the orbital frontal lobe in emotional versus cognitive perspective-taking. *Neuropsychologia, 44*(3), 374–383.

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences of the United States of America, 114*(43), E9145–E9152.

Jacoby, N., & Fedorenko, E. (2018). Discourse-level comprehension engages medial frontal Theory of Mind brain regions even for expository texts. *Language, Cognition and Neuroscience, 0*(0), 1–17.

Jacoby, N., Bruneau, E., Koster-Hale, J., & Saxe, R. (2016). Localizing Pain Matrix and Theory of Mind networks with both verbal and non-verbal stimuli. *NeuroImage, 126*, 39–48.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The Naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences, 20*(8), 589–604.

Jern, A., & Kemp, C. (2014). Reasoning about social choices and social relationships. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.

Jern, A., & Kemp, C. (2015). A decision network account of reasoning about other people's choices. *Cognition, 142*, 12–38.

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition, 168*, 46–64.

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience, 10*(12), 1625–1633.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience, 8*(5), 679–685.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a common-sense moral theory. *Cognition, 167*, 107–123.

Kleiman-Weiner, M., Shaw, A., & Tenenbaum, J. B. (2017). Constructing social preferences from anticipated judgments: When impartial inequity is fair and why? In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Kliemann, D., & Adolphs, R. (2018). The social neuroscience of mentalizing: Challenges and recommendations. *Current Opinion in Psychology, 24*, 1–6.

Koster-Hale, J., Bedny, M., & Saxe, R. (2014). Thinking about seeing: Perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults. *Cognition, 133*(1), 65–78.

Koster-Hale, J., Richardson, H., Velez, N., Asaba, M., Young, L., & Saxe, R. (2017). Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. *NeuroImage, 161*, 9–18.

Koster-Hale, J., & Saxe, R. (2013). Functional neuroimaging of theory of mind. In S. Baron-Cohen, H. Tager-Flusberg, & M. V. Lombardo (Eds.), *Understanding other minds* (pp. 132–163). Oxford: Oxford University Press.

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences of the United States of America, 110*(14), 5648–5653.

Kourtzi, Z., & Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital complex. *Science, 293*(5534), 1506–1509.

Kovera, M. B., Park, R. C., & Penrod, S. D. (1991). Jurors' perceptions of eyewitness and hearsay evidence. *Minnesota Law Review, 76*, 703.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences, 17*(8), 401–412.

Kryven, M., Ullman, T., Cowan, W., & CogSci, J. T. (2016). Outcome or strategy? A Bayesian model of intelligence attribution. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Lafer-Sousa, R., & Conway, B. R. (2013). Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex. *Nature Neuroscience, 16*(12), 1870–1878.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science, 350*(6266), 1332–1338.

Landau, B., & Gleitman, L. R. (1985). *Language and experience: Evidence from the blind child. Cognitive science series* (Vol. 8). Cambridge, MA: Harvard University Press.

Lenci, A., Baroni, M., Cazzolli, G., & Marotta, G. (2013). BLIND: A set of semantic feature norms from the congenitally blind. *Behavior Research Methods, 45*(4), 1218–1233.

Leopold, A., Krueger, F., dal Monte, O., Pardini, M., Pulaski, S. J., Solomon, J., & Grafman, J. (2012). Damage to the left ventromedial prefrontal cortex impacts affective theory of mind. *Social Cognitive and Affective Neuroscience, 7*(8), 871–880.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *The Journal of Neuroscience, 31*(8), 2906–2915.

Lin, N., Yang, X., Li, J., Wang, S., Hua, H., Ma, Y., & Li, X. (2018). Neural correlates of three cognitive processes involved in theory of mind and discourse comprehension. *Cognitive, Affective, & Behavioral Neuroscience, 18*(2), 273–283.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. *The Behavioral and Brain Sciences, 35*(3), 121–143.

Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Wheelwright, S. J., Sadek, S. A., Suckling, J., … Baron-Cohen, S. (2010). Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience, 22*(7), 1623–1635.

Mano, Y., Harada, T., Sugiura, M., Saito, D. N., & Sadato, N. (2009). Perspective-taking as part of narrative comprehension: A functional MRI study. *Neuropsychologia, 47*(3), 813–824.

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology, 62*(1), 103–134.

McNamee, D., & Wolpert, D. M. (2019). Internal models in biological control. *Annual Review of Control, Robotics, and Autonomous Systems, 2*, 339–364.

Miene, P., Borgida, E., & Park, R. (1993). The evaluation of hearsay evidence: A social psychological approach. In *Individual and group decision making: Current issues* (pp. 151–166). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Mills, C. M. (2013). Knowing when to doubt: Developing a critical stance when learning from others. *Developmental Psychology, 49*(3), 404–418.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). General and specific contributions of the medial prefrontal cortex to knowledge about mental states. *NeuroImage, 28*(4), 757–762.

Mitchell, R. E., & Karchmer, M. (2004). Chasing the mythical ten percent. *Sign Language Studies, 4*(2), 138–163.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320*(5880), 1191.

Moeller, M. P., & Schick, B. (2006). Relations between maternal input and theory of mind understanding in deaf children. *Child Development, 77*(3), 751–766.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage, 56*(2), 400–410.

Olson, G. (2003). Reconsidering unreliability: Fallible and untrustworthy narrators. *Narrative, 11*(1), 93–109.

Ong, D. C., Soh, H., Zaki, J., & Goodman, N. D. (2019). Applying probabilistic programming to affective computing. *IEEE Transactions on Affective Computing*, 1.

Ong, D. C., Zaki, J., & Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition, 143*, 141–162.

Op de Beeck, H. P. (2010). Probing the mysterious underpinnings of multi-voxel fMRI analyses. *NeuroImage, 50*(2), 567–571.

Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *Journal of Neuroscience, 30*(30), 10127–10134.

Perner, J., Aichhorn, M., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Thinking of mental and other representations: The roles of left and right temporo-parietal junction. *Social Neuroscience, 1*(3–4), 245–258.

Peterson, C. C., Peterson, J. L., & Webb, J. (2000). Factors influencing the development of a theory of mind in blind children. *British Journal of Developmental Psychology, 18*(3), 431–447.

Peterson, C. C., & Siegal, M. (2016). Representing inner worlds: Theory of mind in autistic, deaf, and normal hearing children. *Psychological Science, 10*(2), 126–129.

Peterson, C. C., & Wellman, H. M. (2018). Longitudinal Theory of Mind (ToM) development from preschool to adolescence with and without ToM delay. *Child Development, 72*(1), 685.

Peterson, C. C., Wellman, H. M., & Liu, D. (2005). Steps in theory-of-mind development for children with deafness or autism. *Child Development, 76*(2), 502–517.

Peterson, C. C., Wellman, H. M., & Slaughter, V. (2012). The mind behind the message: Advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or Asperger Syndrome. *Child Development, 83*(2), 469–485.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America, 98*(2), 676–682.

Richardson, H., Koster-Hale, J., Caselli, N. K., Magid, R. W., Benedict, R., Olson, H., … Saxe, R. (2019). Reduced neural selectivity for mental states in deaf children with delayed exposure to sign language. *PsyArXiv*.

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications, 9*(1), 1027.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology, 4*(3), 328–350.

Said, C. P., Moore, C. D., Engell, A. D., Todorov, A., & Haxby, J. V. (2010). Distributed representations of dynamic facial expressions in the superior temporal sulcus. *Journal of Vision, 10*(5), 11.

Said, C. P., Moore, C. D., Norman, K. A., Haxby, J. V., & Todorov, A. (2010). Graded representations of emotional expressions in the left superior temporal sulcus. *Frontiers in Systems Neuroscience, 4*, 6.

Saxe, R. (2005). Against simulation: The argument from error. *Trends in Cognitive Sciences, 9*(4), 174–179.

Saxe, R., & Houlihan, S. D. (2017). Formalizing emotion concepts within a Bayesian model of theory of mind. *Current Opinion in Psychology, 17*, 15–21.

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science, 17*(8), 692–699.

Saxe, R., & Young, L. (2013). Theory of mind: How brains think about thoughts. In K. Ochsner & S. Kosslyn (Eds.), *The handbook of cognitive neuroscience* (pp. 204–213). Oxford: Oxford University Press.

Saxe, R. R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *NeuroImage, 19*(4), 1835–1842.

Saxe, R. R., Whitfield-Gabrieli, S., Scholz, J., & Pelphrey, K. A. (2009). Brain regions for perceiving and reasoning about other people in school-aged children. *Child Development, 80*(4), 1197–1209.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences, 19*(2), 65–72.

Schick, B., de Villiers, P., de Villiers, J., & Hoffmeister, R. (2007). Language and Theory of Mind: A study of deaf children. *Child Development, 78*(2), 376–396.

Schick, B., & Hoffmeister, R. (2001). ASL skills in deaf children of deaf parents and of hearing parents. In *Society for Research in Child Development International Conference, Minneapolis, MN*.

Scholl, B. J., & Leslie, A. M. (2001). Minds, modules, and meta-analysis. *Child Development, 72*(3), 696–701.

Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E. N., & Saxe, R. (2009). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLoS One, 4*(3), e4869–e4867.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews, 42*, 9–34.

Sebastian, C. L., Fontaine, N. M. G., Bird, G., Blakemore, S.-J., De Brito, S. A., McCrory, E. J. P., & Viding, E. (2012). Neural processing associated with cognitive and affective theory of mind in adolescents and adults. *Social Cognitive and Affective Neuroscience, 7*(1), 53–63.

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science, 15*(3), 436–447.

Shamay-Tsoory, S. G. (2011). The neural bases for empathy. *The Neuroscientist, 17*(1), 18–24.

Shamay-Tsoory, S. G., & Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: A lesion study. *Neuropsychologia, 45*(13), 3054–3067.

Shamay-Tsoory, S. G., Tibi-Elhanany, Y., & Aharon-Peretz, J. (2006). The ventromedial prefrontal cortex is involved in understanding affective but not cognitive theory of mind stories. *Social Neuroscience, 1*(3–4), 149–166.

Shum, M., Kleiman-Weiner, M., Littman, M. L., & Tenenbaum, J. B. (2019). Theory of Minds: Understanding behavior in groups through inverse planning. *arXiv.org*.

Siegal, M., & Peterson, C. C. (1994). Children's theory of mind and the conversational territory of cognitive development. In *Children's early understanding of mind: Origins and development* (pp. 427–455). Hove: Psychology Press.

Skerry, A. E., & Saxe, R. (2014). A common neural code for perceived and inferred emotion. *The Journal of Neuroscience, 34*(48), 15997–16008.

Skerry, A. E., & Saxe, R. (2015). Neural representations of emotion are organized around abstract event features. *Current Biology, 25*(15), 1945–1954.

Sommer, M., Döhnel, K., Sodian, B., Meinhardt, J., Thoermer, C., & Hajak, G. (2007). Neural correlates of true and false belief reasoning. *NeuroImage, 35*(3), 1378–1384.

Spotorno, N., Koun, E., Prado, J., Van Der Henst, J.-B., & Noveck, I. A. (2012). Neural evidence that utterance-processing entails mentalizing: The case of irony. *NeuroImage, 63*(1), 25–39.

Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience, 21*(3), 489–510.

Spunt, R. P., Kemmerer, D., & Adolphs, R. (2015). The neural basis of conceptualizing the same action at different levels of abstraction. *Social Cognitive and Affective Neuroscience, 11*(7), 1141–1151.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences, 22*(3), 201–212.

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences of the United States of America, 113*(1), 194–199.

Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science, 262*(5134), 685–688.

Thornton, M. A., & Tamir, D. I. (2017). Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences of the United States of America, 114*(23), 5982–5987.

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D., & Tenen-baum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (pp. 1874–1882). Red Hook; NY: Curran Associates, Inc.

van Ackeren, M. J., Casasanto, D., Bekkering, H., Hagoort, P., & Rueschemeyer, S.-A. (2012). Pragmatics in action: Indirect requests engage theory of mind areas and the cortical motor network. *Journal of Cognitive Neuroscience, 24*(11), 2237–2247.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., … Zilles, K. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *NeuroImage, 14*(1), 170–181.

Wen, T., Mitchell, D. J., & Duncan, J. (2019). The functional convergence and heterogeneity of social, episodic, and self-referential thought in the default mode network. *bioRxiv, 1*, 753509.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103–128.

Winecoff, A., Clithero, J. A., Carter, R. M., Bergman, S. R., Wang, L., & Huettel, S. A. (2013). Ventromedial prefrontal cortex encodes emotional value. *Journal of Neuroscience, 33*(27), 11032–11039.

Woolfe, T., Want, S. C., & Siegal, M. (2002). Signposts to development: Theory of mind in deaf children. *Child Development, 73*(3), 768–778.

Wu, Y., Baker, C. L., Tenenbaum, J. B., & Schulz, L. E. (2018). Rational inference of beliefs and desires from emotional expressions. *Cognitive Science, 42*(3), 850–884.

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., … Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology, 106*(3), 1125–1165.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of the United States of America, 107*(15), 6753–6758.

Simulation, Predictive Coding, and the Shared World

Robert M. Gordon

The debate between the “simulation” theory and the “theory” theory, initiated in the late 1980s, concerns the source of everyday human competence in predicting and explaining human behavior, including the capacity to ascribe mental states. This competence is approximately what the term *mentalizing* designates, when understood in its broadest sense. Since the 1960s, it was widely assumed that the source of this competence is a body of implicit general knowledge or theory, commonly called “folk psychology” by philosophers and “theory of mind” by psychologists. This was usually understood to consist in a body of general information whose core stipulation is that intentional action is a causal product of the agents’ beliefs and desires.

The “simulation” theory locates the main source of mentalizing competence in a procedure or set of procedures called “simulation,” or “mental simulation.” Introduced by philosophers (Goldman, 1989; Gordon, 1986; Heal, 1986), this account is usually thought to challenge the very assumption that mentalizing is an application of an implicit theory of mental states. The “theory versus simulation debate” soon became a topic of interest among developmental psychologists and later received attention in linguistics, social cognitive neuroscience, and social robotics.

One of the initial motivations for a simulation account of mentalizing was that it seemed to spare the brain the overhead costs predicated by the prevailing “theory” theory. These were the costs of acquiring, storing, and utilizing the theory. An important part of the simulationist response was to ask why a system would need to invest in a general theory or model of *systems like itself*. Wouldn’t it be more economical simply to use itself as a stand-in for these other, similar systems? Of course,

R. M. Gordon (✉)
Department of Philosophy, University of Missouri, St. Louis, St. Louis, MO, USA

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_12

"using itself as a stand-in" may be understood in different ways, and, as a consequence, various distinct simulationist approaches have developed.

The approach presented here aims to show how one's own action planning system may serve as a stand-in, in a sense that will become clear, for the action planning systems of other agents. Building on recent work on inverse planning, it explains how mentalizing by simulation can offer vastly greater economies than the mere elimination of the information-rich overhead required by a "theory" theory. The simulation approach, as presented here, is in fact much in line with the current view in psychology and neuroscience that neural systems tend to reduce metabolic and other expenses by conforming to a predictive coding strategy. This is a strategy of "guessing ahead." Rather than waiting for the world to bombard us with new information, the system makes its latest best guess as to what will be coming in. This process of predicting input values minimizes the need for new information input, in that only discrepancies, or information that conflicts with the predicted values (prediction errors), need be encoded.

Indeed, simulation has been compared to compression schemes commonly used in the digital transmission and storage of video content (Gordon, 1992). These schemes exploit the likelihood that video content will be redundant in a number of ways. Most important is temporal redundancy. Typically, little or no visual content changes in, say, the 30th of a second that separates one frame from the next; successive frames in a video sequence are nearly always very similar. Therefore, it is an efficient strategy to treat each frame as "predicting" its successor. The default, or uncorrected, prediction would yield a sequence of undifferentiated frames: essentially, a still picture. Any corrections, or departures from this default, are likely to be relatively small, requiring minimal resources to encode these differences.

Video compression was an early engineering application of predictive coding. A comparable simulation account should show how our mentalizing system exploits massive redundancies to achieve extreme code compression and resource parsimony. Simulation, as I understand it, does just that, I believe.

Two Kinds of Projection

In broad view, simulation exemplifies a type of predictive strategy that begins with what is in effect a forward model—projection—creating a default expectation. However, I should note that the term *projection* (or *self*-projection) is ambiguous in this context, yielding two distinct metaphors: "projection *onto*" and "projection *into*." To project *onto* another person or entity is to push or impose (etymologically, *to throw*) one's own image (or perspective or "way of seeing things") onto the other, thereby assimilating the other to oneself.[1] Projection *into*, on the other hand, is

[1] To the Freudians, it is a defence mechanism by which one deviously assigns to someone else a mental state or trait one cannot accept in oneself.

metaphorically a kind of travel, where it is we who move, rather than our image: we are transported into a perspective that is not currently our own. Understood in this second way, projection is not an imposition of one's own perspective, but rather a shift to a different perspective: This may be the perspective of another human or other sentient being. Or, it may involve mental time travel to a past perspective (as in episodic memory) or a future perspective (prospection). It might be travel to a counterfactual perspective, to an "alternative" past, present, or future. It has been suggested that the various forms of projection *into* may in fact be supported by a single specialized brain network (Buckner & Carroll, 2007).

On the projection-and-correction account, simulation entails both kinds of projection. It begins with projection *onto* another presumably sentient being, imposing one's own perspective as an initial *a priori* prediction of the other's, and it ends at, or at least aims at, being projected *into* the other's perspective. Simulation proceeds from *onto* to *into* by a series of "corrections," or *corrected projections onto*. These corrections may come from a comparison of predicted behavior with observed behavior, from internal resonances to observed behavior, from contextual evidence of various kinds, and other sources. In short, mentalizing by simulation begins with an uncorrected projection *onto* a target and then, in response to predictive errors, tests hypothetical modifications of this projection until a good enough projection *into* is achieved.

An Important Problem Not Addressed Here

The projection-and-correction account of simulation theory conforms to what Andy Clark calls "the core predictive coding strategy." However, it does not entail a much more ambitious package, which Clark distinguishes as "hierarchical predictive processing" (Clark, 2013). Not only does the latter analyze neural inference as a process of prediction and correction, but it also aims to specify the inference mechanism by which predictions are made and corrected. It posits a hierarchy of intermediate predictions and corrections, each of which operates by Bayesian inference. In this manner, higher-level predictions are thought to pass stepwise down to lower levels, and lower-level corrections are in a similar fashion passed back up to higher levels. This has been a very influential idea (Clark, 2013; Friston, 2005; Rao & Ballard, 1999). However, I believe that the argument that follows does not require commitment to the hierarchical Bayesian account of prediction and correction. The focus here will be on the initial projections, or the a priori starting point, or what I will be calling the default condition for mentalizing.[2]

[2] The temporal terms "initial," and "starting point" should not be taken to imply that each instance of analysis by inverse planning begins with uncorrected agent-neutral coding.

Inverse Planning

The holy grail for a theory of mentalizing is to account for our capacity to grasp the intentions behind observed behavior; beyond the intentions, detect the goals and reasons motivating these intentions. The aim, one might say, is to illuminate the background that makes the observed behavior unsurprising.

To address the question of how the brain interprets the observed actions of others, it has been suggested that we adopt a predictive framework that has been particularly fruitful in studies of vision:

> it is often said that "vision is inverse graphics" - the inversion of a causal physical process of scene formation (Baker, Tenenbaum, & Saxe, 2006).

Just as the interpretation of a visual scene might involve, essentially, using in reverse the process of producing such a scene, so the interpretation of another's behavior might be understood as a comparable inverse problem (Baker, Saxe, & Tenenbaum, 2009, 2011).

> By analogy, our analysis of intentional reasoning might be called "inverse planning", where the observer infers an agent's intentions, given observations of the agent's behavior, by inverting a model of how intentions cause behavior. (Baker et al., 2011)

The process is *inverted* in that, instead of proceeding forward from a given intention to its behavioral execution, it takes the behavior as the given and determines the intention most likely to have produced it. The planning process would thus be used as a mechanism for testing hypotheses about underlying intentions.

In the broadest terms, inverse planning exemplifies hypothesis-testing as unconscious inference, an idea introduced in the perceptual realm by Helmholtz (1856). The proposal bears some resemblance to "hypothetico-practical" inference (Gordon, 1986), modeled on a traditional model of the scientific method, hypothetico-deductive inference. Instead of forming hypotheses and *deducing* consequences that match observations, hypothetico-*practical* inference would form hypotheses and then *act on* them, *producing* consequences that match the observed behavior of the other agent. (Typically, the resulting action would be merely covert, inhibited from any outward expression that might be perceptible to others.) This was conceived as an inference requiring personal agency—as something *I* do, rather than as computational operations of a "subpersonal" neural system such as an action planning system. This chapter concerns such computational operations, building on a variant of the inverse planning proposal.

Model or Reuse?

The remainder of this chapter builds on a variant of the inverse planning thesis proposed by Baker et al. (2009) and Baker et al. (2011).

The term "inverse planning" appears to suggest that the very mechanism that is used to plan our own behavior may be reused as a platform for testing hypothetical explanations of the observed behavior of other agents. However, Baker et al. (2011) actually propose something more complicated. The authors speak of inverting a *model* or *theory* of the planning process. One might understand them to be saying that the action planning system itself can be used as a general model of an action planning system. But this is not their view. As they point out,

> on first glance our work appears most consistent with the "theory-based" approach. Formalizing an intuitive theory of mind was in fact one of our original motivations …. On a theory-based interpretation, inverse planning consists of inverting a causal theory of rational action to arrive at a set of goals that could have generated the observed behavior, and inferring individual goals based on prior knowledge of the kinds of goals the observed agent prefers. (A closely related view [Jara-Ettinger, 2019] treats action understanding as inverse reinforcement learning: determining what model of the world and what positive and negative reinforcers would best explain an agent's observed actions.)

The theory-based approach attributes to the brain a capacity for detachment: it *stands back from its own operations* and employs instead a general theory or model of these operations. As distinct from actual action planning, the theory theorist proposal is that in mentalizing about others the brain engages in *plan-theorizing, theorizing about* the steps in the other's planning process. The proposal assumes that we humans have an intuitive theory of mind and that our brains employ this theory not only in our explicit attributions of mental states but also in its unconscious subpersonal neural processing. I will call this *inverse plan-theorizing*. Thus understood, it does not make use of our capacity for planning: it is not inverse planning as such, i.e., an inverse reuse of one's own action planning system.

Although Baker et al. (2009) explicitly develop this plan-theorizing version of inverse planning, they acknowledge that a simulation-based account would cover the data just as well as their theory-based account. The simulation account would use, not a *model* of the planning process, but the planning process itself (running offline), as a mechanism for testing hypotheses about underlying intentions:

> On a simulation account, goal inference is performed by inverting one's own planning process - the planning mechanism used in model-based reinforcement learning - to infer the goals most likely to have generated another agent's observed behavior.

If indeed such reuse of its own "first-person" planning system (what I will call *1p* inverse planning) would be sufficient for goal inference, the question arises, Why would the brain need to operate instead on a model of the planning process? Here again, using an existing system would avoid the overhead costs of storing and utilizing an information-rich theory or model.

Moreover, first-person inverse planning would seem to be the proper analog of the inverse graphics account of vision. As inverse graphics is "the inversion of a causal physical process of scene formation" (Baker et al., 2011), so inverse planning should be the inversion of *a physical process* of action determination—*not* the inversion of a causal *theory of* a physical process of action determination. The "vision is inverse graphics" idea is generally understood to be an analysis-by-synthesis paradigm, and analysis by synthesis is not analysis by a *theory of* synthesis.

> The perceptual system … is a mechanism for the hypothetical "synthesis" of natural images, in the style of computer graphics. Perception (or "analysis") is then the search for or inference to the best explanation of an observed image in terms of this synthesis. (Yildirim et al., 2020)

In other words, just as visual perception is thought to weigh alternative hypothetical ways of building a given scene, understanding action in terms of goals and intentions would be a search among alternative hypothetical ways of generating (planning) a given action, in an effort to find the most plausible simulation of the planning that might have generated the action. Action understanding as inverse planning would thus really be, in analogy to vision as inverse graphics, a case of analysis by synthesis.

In addition, inverse planning, as a reuse of one's own planning system, would be in a position to exploit responses to the behavior of others that are themselves reuses of one's own motor system. These would include various forms of motor resonance, including mirror neuron activation, motor mimicry, and suppressed action imitation. Rather than having to work with bare visual input, first-person inverse planning could work on lower-level input already formatted in terms of first-person motor planning. This suggests a considerable advantage not available to a model-based understanding of inverse planning.

Use and Reuse of Action Planning

I will suppose then that the human action planning system has, in addition to its primary use in generating one's own actions, a reuse, or secondary use, in which the planning process is inverted in order to infer the goals and reasons that lie behind another agent's observed behavior. This dual use of the same system, I will argue, offers two major advantages. First, it would provide an important head start in understanding the basis of another's behavior, and, second, it would make possible the most economical coding available to a mentalizing system.

It appears likely that the secondary use of the action planning system, namely, inverse reuse for explanatory purposes, runs concurrently with its primary use, for generating one's own actions; otherwise, we would have to suspend our own actions in order to understand the actions of others. Thus, the system is translating existing inputs into action and at the same time looking for hypothetical inputs that would explain the perceived actions of others. Concurrent processing for self-action and other-understanding would be consistent with evidence of "motor contagion," or interference effects between observed and executed actions. First noted in the case of biological movements, it has been suggested that motor contagion may be "the first step in a more sophisticated predictive system that allows us to infer goals from the observation of actions" (Blakemore & Frith, 2005). Indeed, recent research indicates that such interference is markedly increased when the observed movement is directed toward a visible goal (Bouquet, Shipley, Capa, & Marshall, 2011). This interference suggests a competition for resources, and thus that the same, or strongly overlapping, neural resources are employed concurrently in goal-directed action planning and in interpreting the goal-directed actions of others.

Such concurrent double employment raises the question: What, if anything, must *change* as the planning system switches from primary use to reuse, and from self to other? Specifically, what happens to the existing inputs? When the system switches to inverse planning as it seeks to explain another's behavior, does it clear the slate and approach the task with no a priori top-down commitments? More specifically, does it suspend the beliefs, desires, preferences, emotional valences, affordances, and other influences on one's own action planning?

Consider three options:

1. The *suspend-all* option: Suspend all existing inputs and start with a blank slate. The mental states that lead to one's own actions have no informational value for understanding the underlying causes of another agent's behavior.
2. The *keep-all* option: Keep all existing inputs, add no others, and seek the best explanation of the other agent's behavior strictly on the basis of one's own mental states.
3. The *modify-as-needed* option: Keep existing inputs, but allow them to be suspended or modified as needed, and allow new inputs as needed.

Options 1 and 2 should be rejected out of hand. Consider this example: we see a puddle in someone's path, and we expect the individual to deviate from a straight path. When they do, we readily surmise that they did so for a reason, namely, that there was a puddle in their path. More fully, they did so because stepping in a puddle gets you wet, and so they deviated in order to continue on their journey without getting wet.

Option 1, the suspend-all option, might make sense if it were useless, or at least a bad bet, to project onto the other our own perception of a puddle, as distinct from a patch of dry pavement or, for that matter, a manhole or a bed of geraniums. Likewise, if it were useless to project our own desire to avoid getting wet under similar circumstances. However, such projections, and the expectations and explanations based on them, are not bad bets in general, even if they are sometimes in need of correction.

Option 2, the keep-all option, lies at the opposite extreme. It locks all explanations of the behavior of others into our own mental mold, leaving us unable to accommodate differences between ourselves and others. It does not allow inverse planning to move beyond simple projection onto others.

Puddle-avoidance may seem to present a trivial problem of action understanding. It is common behavior, and it appears to be a matter of common sense. Instead of calling on our own desire to avoid puddles, we might simply apply the generalization, "People tend to avoid puddles." Similar generalizations would apply to avoidance of snarling dogs, bears, and "shady-looking" people. Such generalizations do not necessarily compete with projection, however; in fact, it is at least plausible that they are themselves products of projection. I am fairly confident that my own acceptance of a generalization like "People tend to avoid stepping in puddles" is not based on extensive observation of people confronted with puddles. More likely, it is a projection of my own desire. Further evidence emerges in cases of conflicting desires. For example, will a runner of 100 m run around a puddle on the track, losing time, or plow through it? There is no general rule I apply to this question, certainly not one based on observation. Rather, I project my own competing desires, letting the specific situation (e.g., is this a practice run or the real thing?) dictate the answer.

Agent-Neutral Coding

Although such projection would provide inverse planning with an important head start, certainly better than to begin the process without any a priori predictive input, a capacity to modify the default projection would be advantageous as well. This leaves us with Option 3, the modify-as-needed option. As in Option 2, inverse planning (i.e., hypothetically planning) of another's actions starts with the existing inputs to planning our own actions. I will call this *agent-neutral* coding. The same undifferentiated coding would serve in two capacities, as our own desire to avoid puddles in our path and, within the context of inverse planning, as the other's desire to avoid puddles in their path. However, per Option 3, the agent-neutral coding is subject to revision.

By *agent-neutral* coding I mean *identical*, *undifferentiated* coding, the same for self and other. This may seem problematic. Surely, the brain must be able somehow to distinguish its own states from the represented states of others. Authors who speak of *shared representation*s, or *shared self-other representation*s, usually emphasize that coding for self and other is overlapping but not identical (Decety & Sommerville, 2003). Jeannerod and Pacherie (2004) have argued that, when the actions of others are simulated in the brain by representations shared with similar actions of our own, they elicit agent-neutral or unattributed ("naked") intentions.[3] These neutrally coded intentions leave open the question, "Whose intention is this?" To determine who the author of the intention is requires collateral information. Because such information is fallible, misattribution of intentions is possible—and is in fact often exhibited in people with schizophrenia.

One might think the same would be true for the beliefs, desires, and emotions that provide input to action planning. It is my own mental states that provide input to the forward planning of my own actions, and representations of the other's mental states that feed into the inverse use of the planning system to explain the other's behavior. It might be supposed that the system has to distinguish these in some way. But this is not so. Unlike intentions and motor plans, beliefs may remain happily

[3] Joëlle Proust and Shaun Gallagher independently called my attention to similarities between my agent-neutral coding and the agent-neutrality of intentions posited by Jeannerod and Pacherie (2004).

undifferentiated, and failure to differentiate is not only not pathological, but it is also the norm. What the system needs to "know" is, simply, that there is a puddle in the path; it can deal with undifferentiated, impersonal "facts," without marking them as facts-to-me, facts-to-you, or facts-to-another—or, in other words, as facts *as I believe them to be*, or you, or another. Moreover, as will be argued, simple "factive" explanations, such as "She stepped to the side because there was a puddle in the path," are the preferred form of action explanation, in contrast to "because she believed …" explanations. (Use of "because she believed …" is taken to imply that there was reason not to use the simple factive form.)

Coding for beliefs would *begin* as agent-neutral, in the sense that any differentiation would be the result of intervention of some sort: identical coding for self and other would be the default. Would the same would be true of coding for desires and emotions? Regarding emotions, it is important to distinguish our own emotions from our empathetic responses to another's emotion, and maybe pathological not to. Nevertheless, it is common to think of elements of the environment as disgusting, frightening, and so forth, without specifying "to whom?" Likewise, the world may be seen as motivationally charged, or valenced. Objects may be seen as attractive or repulsive, without an implicit "*to* (somebody)." Even possible future states of the world may be regarded as emotionally and motivationally charged in this nonrelative way. This would suggest that undifferentiated agent-neutral coding would present us with a shared world of facts and emotive and motivational valences—the rich shared world that appears to us, I will argue, as a consequence of maximal code compression.

Summary So Far

The argument so far centers on three claims

1. Understanding the intentions behind actions is accomplished by inverse planning.
2. The 1p (simulation) version of inverse planning is correct.
3. The default top-down inputs to 1p inverse planning are in agent-neutral coding.

If 1–3 are right, then inverse planning gets a free head start, which can then be corrected as needed (Option 3 above).

If 3 is right, then inverse planning defaults to the greatest possible code compression. In the default condition, inverse planning requires no new input coding to explain the actions of others.

Projection by Default

If we understand simulation in terms of default agent-neutral coding, then we have to reject a well-known account of the simulation theory: that it requires introspective recognition of one's own (actual or pretend) mental states (metacognition), followed by attribution of the same states to the other individual (Goldman, 2006). Agent-neutral coding clearly would support a more economical account of simulation, one that requires neither metacognition nor self-other inference (Gordon, 1995, 1996). It is simply by default that the inputs to inverse planning are the same as the inputs to forward self-planning; This carryover is not established through an inferential leap from self to other, but rather, as I suggested, simply by omission: that is, crossing the self-other border without doing anything to *alter* the existing inputs.

It is often supposed that attributions of beliefs and desires constitute the very heart of our everyday understanding of the behavior of others. This questionable assumption is built into the use of terms such as *mentalizing*, *theory of mind*, *mind-reading*, and *folk psychology.* These expressions appear to suggest that our understanding of others is based on learning what is going on in their minds, particularly the mental states and processes that cause their behavior. I think this places undue emphasis on attributions to the individual, as opposed to attributions to the situation, or, more broadly, the world. Our everyday effort to make sense of the behavior of others is chiefly an attempt to discern the reasons for their actions, and to discern these reasons is, in general, to know *what it is about the world* that explains their actions. It is not, in general, to discern the *state of mind* behind the action. There are several reasons for asserting the primacy of the world in our understanding of others.

First, it makes evolutionary sense that people would prefer explanations of action and emotion that look to the world, rather than to mental states, such as beliefs about the world. It is often useful to identify items in the common environment, especially threats and rewards, and to explain behavior in terms of facts about them. We want to know what it is about the world that is making someone run: perhaps something behind them (from which we should run as well) or something ahead of them (to which we might want to run as well). The parent wants to know what it is about the environment that frightens or upsets the child; in social referencing, the child wants to know what the caregiver is responding to so that it can copy and learn the response.

Second, linguistically, reference to the beliefs of the agent is generally treated as a fallback. "Why are you stepping off the path?" Ordinarily, we wouldn't respond, "Because I think (or: believe) there is a puddle." Rather, we say, "Because there is a puddle." Likewise, in the third person, "Why did he step off the path?" Mentioning what the agent thought or believed would imply that there is something wrong with a simpler explanation in terms of "the fact" that there is a puddle. For example, we worry that the agent may have been tricked by an illusion. Our explanations, whether of our own actions or the actions of others, default to the factive. (The relevant facts, I should add, may be facts *about* mental states: for example, "I'm calling the dentist because I have an awful toothache," and, "I'm eating now because I'm hungry [or: bored].")

Third, the most economical strategy for mentalizing, other things being equal, would be one that minimizes individuation, or information tagged to specific individuals. That is, it would minimize the need for explicit mentalizing, in the sense of judgments about mental states or processes. In the default case, with uncorrected agent-neutral coding, the actions of others would be interpreted in terms of a shared world—that is, to the world on the basis of which we ourselves act. Mentalizing, on this account, would be called on to complement or to correct what is passed along through agent-neutral coding. It would be reserved for cases in which a shared world proves inadequate to predict or explain the actions or emotions of particular individuals.

The economizing extends beyond the preference for facts over the beliefs of individuals. It is a feature of our phenomenology that we see objects and events as having, among other properties, *emotive* qualities: the qualities of being scary, repulsive, attractive, embarrassing, shameful, pleasing, and so forth. Such "externalizations" have the virtue of setting expectations and making the corresponding responses by others—being scared, repelled, attracted, embarrassed, etc.—unsurprising. They can limit the need for coding to the exceptions, the surprising outliers. Similar considerations hold for the affordances of objects, making their standard uses unsurprising and reserving special coding for surprising, nonstandard uses, misuses, and nonuses. Phenomenologically, these properties of objects would be carried over as we slip seamlessly from the forward planning of our own actions to our hypothetical planning of the other's actions. It seems obvious that the more the brain is able to place the causes of actions and emotions in an objective world, the less coding it will need.

Perspective-Taking and Positional Correction

Probably the most familiar type of correction is spatial perspective-taking. For example, to a stranger observing the scene from a distance, the bear now approaching me is not likely to feel threatening, or in any case as threatening as it does to me. The threatening (or nonthreatening) emotive quality of the bear may be seen as a function of one's location relative to the bear—or, the bear's location and vector in egocentric space. With the ability to move mentally into another's spatial perspective, individual differences become mere positional differences. That is, it is a good starting bet that (unless there is evidence to the contrary) any individual in the same position will see the bear as threatening. With the operation of "putting ourselves in the other's place" by spatial perspective-taking, we are able to restore the economic advantages of a shared world. We allow the threatening quality to remain out there in the bear, or rather in the bear-from-a-point of view. We need not represent it as a function of individual mental makeup, even if some individuals may be found immune to the standard bear-approaching-me response.

Although it is spatial perspective-taking that gives us the general metaphor of "perspective-taking," "adopting the other's point of view," and "putting ourselves in the other's place," many other kinds of corrections may be considered broadly perspectival, or positional. For example, differences in social or occupational role may be bridged by a kind of perspective shift: student/teacher, worker/manager, diner/waiter, patient/doctor, and consumer/salesperson. In these cases, as in differences in spatial perspective, it may be sufficient to shift to a generic "point of view," or, as we say, to understand where the other is "coming from," to explain the other's actions, without explicit mentalizing. That is, it may be a good starting assumption that a person in a given "position" will act in more or less the same "standard" way, an assumption that may underlie the notion of generic "scripts" of action sequences postulated by Schank and Abelson (1977). Such an assumption would exploit positional redundancies and limit new input to deviations from the standard.

Caveats and Qualifications

1. It seems unlikely that each instance of analysis by inverse planning would start with unrevised agent-neutral coding, and thus uncorrected projection onto the other. With experience and maturity, we should be able to make adjustments before observing the other's actual behavior. We should be able to adjust beforehand to the other's spatial perspective or, more generally, the other's epistemic situation. Likewise, known personal history, social and institutional roles, relationships, and culture may pre-adjust our projections away from uncorrected projection onto the other. Our expectations may conform to templates for particular individuals or classes of individuals, as I explain in a later section; these templates may be shaped by irrational biases as well as by evidence.
2. Some mental states do not as a rule cross from self to other. Our pains, for example, should be left behind: generally, it would not be a good bet to project onto others the physical pain we feel; likewise, physical pleasure, hunger, and bodily sensations. These, of course, influence our own actions, but they are not generally allowed to motivate others in the same ways. Perhaps, these are simply excluded from agent-neutral coding; or, as I think more plausible, infants, in the process of developing self-awareness and self-differentiation, acquire "export prohibitions" that, so to speak, prevent these inputs to action planning at the border between self and other.

An Evolutionary Perspective

Agent-neutral coding may be expected to present problems for understanding the behavior of people of very different cultures inhabiting far-away lands. We may seem to share only the bare physical parameters of earthly human life: we breathe, we eat, we sleep, we procreate; the sky is blue (more or less) and grass is green (more or less). Undiscriminating projection onto such people might seem so wide of the mark that, without extensive correction, we could neither predict nor explain much of their behavior.

And yet it is easy to forget that until the very recent past, nearly all human social encounters would have occurred among people in close cultural as well as physical proximity. For much of the history of our species, people would have had little need to depart from the simple strategy of looking to the shared world, with its facts, emotive qualities, affordances, attractions, and repulsions, for the causes of observed behavior. With like-minded individuals in close spatiotemporal proximity, people would have gotten by with few corrections beyond the export prohibitions and positional adjustments such as those mentioned earlier.

The small social groups in which early homo sapiens lived would have shared a local environment and probably formed similar mental maps of that environment. They would have had a shared understanding of the elements of their environment and of the causal properties of these elements, as well as their affordances and emotive qualities. There would have been wide agreement on which elements were salient, menacing, frightening, attractive, or disgusting.

Of course, even in the distant past, there would have been cracks in the vault of this shared world. Adjustments would have to be made to differences in temperament, in sensory and cognitive capacities, in knowledge, acculturation, and in goals. Such differences would of course have been salient and noteworthy against the more or less fixed and predictable shared background. But they would have been relatively rare in social groups with strictly limited horizons.

We should note that a social predictive system doesn't just exploit redundancies; it reinforces them and also adds new redundancies. To benefit from the redundancies within our small group, it would help to have grown up within the group. For much of the redundancy within the group is likely to have been a product of earlier corrections. This is especially true of infants and young children, who tend to fill in or replace their own view of the world with those of their adult caregivers. For example, in social referencing, the child observes the adult's response to *x* (a person, object, or situation), and then copies the response. For example, if the adult appears frightened by *x,* the child will be frightened by *x*. This not only modifies any prior expectation the child may have had of the adult's response to *x*; it also modifies the child's future responses to *x*. The child's response now conforms to the adult's. This increases redundancy within the group and makes future predictions easier.

Ignorance and False Belief

I argued earlier that our explanations, whether of our own actions or the actions of others, default to the factive. Mentioning what an agent thought or believed would imply that an explanation in terms of the corresponding "fact" would in some way be defective. For example, we observe someone in broad daylight walking nonchalantly into a deep puddle (or: a lamppost). Why? What accounts for his aberrant behavior? Answer: he was looking at his cell phone, oblivious. We could have predicted it, and now we can explain it.

We make his behavior unsurprising by disconnecting or "decoupling" the fact that there was a puddle in his path from the input to inverse planning. Decoupling a fact from inverse planning is a way of marking ignorance of fact. Ignorance in turn may engender false belief: because he was ignorant of the fact that there is a puddle—out of touch with the facts concerning his current environment—he continued operating on the false default assumption of a puddle-free path.

In the classic false-belief experiments (Wimmer & Perner, 1983), the task is one of *predicting* behavior, rather than interpreting or explaining a given action. One's action planning system may be recruited for this task, but it would be through a vicarious *forward* use of the system, rather than an inverse use. As before, you would decouple the system from the events that transpired in the other's absence. Absence from a scene creates a blind patch, a scotoma of ignorance. Given ignorance, the forward use of the planning system would predict that other would not do the well-informed or "correct" thing (see also Perner & Roessler, 2010).

Agent-neutral coding and the possibility of toggling between knowledge and ignorance would give us the neural underpinnings for two theses long held by Josef Perner: first, that well before they have an explicit grasp of belief attribution, young children are quite capable of explaining action in terms of the external situation, and second, that older children and adults use the same type of explanation young children use, except in the occasional cases where it proves inadequate; then they must fall back on explanations that mention the mental states, especially the beliefs, of the agent. Young children and, where possible, older children and adults

> make sense of intentional actions in terms of justifying reasons provided by "worldly" facts (not by mental states). (Perner & Roessler, 2010)

The young child's conception is all we usually call upon, because it is typically all we need. This comes to saying that explaining and predicting actions in terms of actual situations or facts is our default mode of explanation and prediction, the mode we employ unless we find some reason not to. Only where this appears inadequate do we invoke beliefs in our explanation.

Anchoring and Adjustment

The projection-and-correction understanding of simulation bears a close kinship to recent work based on the "anchoring and adjustment heuristic" originally proposed by Tversky and Kahneman (1974) (see Epley et al., 2004; Tamir & Mitchell, 2010). The idea is succinctly expressed in Epley et al. (2004):

> people adopt others' perspectives by initially anchoring on their own perspective and only subsequently, serially, and effortfully accounting for differences between themselves and others until a plausible estimate is reached.

The serial adjustments are conceived as a process of moving out from one's own perspective, as necessary, to more "distant" perspectives, with hypothesis-testing at each stop along the way—a process that should take longer, the farther out one goes.

Tamir and Mitchell (2010) cites evidence that

> the MPFC [medial prefrontal cortex] subserves the use of self-projective simulation as one route to understanding other minds.

They suggest that

> subregions of the MPFC not only may use the self as an anchor point from which to understand others but also may actively allow perceivers to adjust their inferences about another person.

That is, the MPFC may make possible, not only self-projective simulation, but also the corrections or adjustments that enable perceivers to accommodate minds different from their own.

According to the view developed in this paper, the default to an "anchor" is a consequence of predictive coding at the neural level, specifically, agent-neutral coding in self-planning and inverse other-planning. Indeed, Koster-Hale and Saxe (2013) found anchoring-and-adjustment to be implemented by predictive coding in the medial prefrontal cortex (MPFC):

> When specific information about a person's reputation or traits is unavailable, we may predict others' preferences by assuming that they will share our own preferences (Krueger and Clement, 1994; Ross et al., 1977). In one series of studies (Tamir and Mitchell, 2010), participants judged the likely preferences of strangers (e.g., is this person likely to "fear speaking in public" or "enjoy winter sports"?) about whom they had almost no background information. Under those circumstances, the response of the MPFC was predicted by the discrepancy between the attributions to the target and the participant's own preference for the same items: the more another person was perceived as different from the self, for a specific item, the larger the response in MPFC.

This suggests that in at least one area of mentalizing, the default is uncorrected projection onto a target, with adjustments in response to contrary information. Essentially, subjects respond initially as if the question were about themselves and then make adjustments for differences.

The only caveat I would raise concerns the explanation of the MPFC responses: that they reflect the subject's "assuming that they will share our own preferences." This imposes on the data the unwarranted assumption that the default to an anchor is an optional belief-based heuristic, a shortcut we use because it seems like a good idea. If the argument of this chapter is correct, the default to an anchor is a structural feature of 1p inverse planning, a consequence of doing nothing to alter the input to self-planning—including, of course, our preferences.

There is evidence that the adjustments made in these anchoring-and-adjustment experiments do not consist in theorizing about personalities different from one's own, but rather in impersonating them. The subject becomes a shapeshifter, modifying her brain to respond in character. The amygdala, in particular, is an active participant in this impersonation. It plays a role when, as in several of the anchoring-and-adjustment studies, we predict the responses of a (hypothetical) person with an emotional disposition that differs from our own (Gilead et al., 2016). It is also activated when we respond empathetically to another's emotional suffering (Bruneau et al., 2015). I think it is reasonable to speculate that similar changes occur when we try to anticipate the behavior of familiar individuals or try to understand the intentions behind their actions. Our actual mentalizing in such cases might begin with a priori templates, comparable to the ready-to-use set of transformations of an actress getting into an accustomed role: her Lady Macbeth template, her Blanche

DuBois template, and so forth. Something comparable may be true of our simulation of particular individuals or classes of individuals. We project with a fixed set of adjustments, and these become *a priori* expectations.

Conclusion

This chapter develops the idea of simulation as a predictive strategy for mentalizing. The predictive aspect consists in an initial projection onto the other, which is then corrected and revised as needed. Such a projection is implemented by allowing our own mental states to govern, not just our own behavior, but also our hypothetical interpretations of the observed behavior of other agents. The mechanism proposed for this is a variant of the inverse planning mechanism put forward by Baker et al. (2009). According to what I call *first-person* (or *1p*) inverse planning, our own action planning system is recruited as a hypothesis-testing device. In the default condition, the top-down inputs to interpreting the behavior of other agents would simply be the same as the inputs to planning our own actions. This agent-neutral coding would be modified as needed to generate intentional actions that come close enough to matching bottom-up perceptual-motor input from observing the behavior of others. In creating and evaluating alternative reconstructions of the processes that gave rise to the other's behavior, the system performs an analysis by synthesis, like the inverse graphics account of visual perception.

In reconstructing the processes behind the other's action, inverse planning locates the agent's reason or reasons for acting, as far as possible, within a shared world of facts, affordances, emotive and motivational valences, and other "objective" properties. Where this is problematic, as in the case of ignorance, false belief, and nonstandard emotional responses, inverse planning locates the causes within the mental states of the individual agent. Shared world explanations have a number of advantages over those requiring explicit mentalizing: they can identify environmental threats and rewards, they are conceptually and linguistically less demanding, and they achieve greater code compression. If this is correct, then we must reject the common assumption that explicit mentalizing, or mental state attribution, is the paramount explanatory aim of the procedures we lump under the term *mentalizing.* The aim is rather to interpret behavior in terms of a shared world where this is possible and to diagnose cases where it is not.

The notion of a shared world may seem quaint today, when few people belong to small, culturally and geographically isolated groups. Living successfully in the world of social media, especially, would seem to demand constant sophisticated and resource-hungry mentalizing to accommodate disparate voices. And yet, notoriously, we seem to manage by stepping into echo chambers of like-minded individuals creating the semblance of a shared world. These "pocket worlds" may reflect a more general human tendency to form limited "social niches" with mutual expectations based on shared "cultural affordances" constituting a shared world (Veissière, Constant, Ramstead, Friston, & Kirmayer, 2019).

The Shared World Comes First

I said at the beginning that the simulation theory stipulates that the main source of everyday human competence in predicting and explaining human behavior consists in a certain procedure or procedures. Broadly speaking, simulation allows us to put ourselves in the other's place when we are not, in the relevant sense, already in the other's place.

However, in the light of what I have argued here, I would have to say that simulation thus understood is not actually the main source of this competence.[4] Rather, the main source is not a procedure at all. It is the persistence, the non-alteration, of top-down inputs as our mechanisms for decision-making and planning switch to inverse use for testing hypothetical explanations of the actions of others. This identity, or agent-neutral coding, of inputs has the effect of projecting onto others a shared world within which we act and interact. Strictly speaking, the brain doesn't have to do anything to accomplish this; rather, what is required is inaction: simply leaving the inputs unchanged as it switches from self to other. The shared world is the default.

The fact that such a system of action interpretation, starting with no additional expenditure of resources and adding only as needed, would be as cost-effective as a system can be is reason for confidence that it is the system we have.

Acknowledgments The author is enormously grateful to Michael Gilead for guidance, encouragement, and feedback at several steps along the way. Thanks to Gualtiero Piccinini and Frank Van Overwalle for helpful early discussions of the topic. The idea of simulation as neural reuse owes much to the influence of Susan Hurley. Special thanks to Karl Friston and Timothy Williamson for very welcome comments and suggestions for future development.

References

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour, 1*, 1–10.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*, 329–349.

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society* (pp. 2469–2474).

Baker, C. L., Tenenbaum, J. B., & Saxe, R. R. (2006). Bayesian models of human action understanding. In *Advances in neural information processing systems* (Vol. 18, pp. 99–106). Cambridge, MA: MIT Press.

Blakemore, S.-J., & Frith, C. (2005). The role of motor contagion in the prediction of action. *Neuropsychologia, 43*, 260–267.

[4] Daniel Dennett once pointed out to me that if simulation were our regular way of understanding others, we wouldn't have to tell anyone, "Put yourself in their place." Regarding "putting in place" simulation, I believe he was right.

Bouquet, C. A., Shipley, T. F., Capa, R. L., & Marshall, P. J. (2011). Motor contagion: Goal-directed actions are more contagious than non-goal-directed actions. *Experimental Psychology, 58*(1), 71–78.

Bruneau, E.G., Jacoby, N., Saxe, R. (2015) Empathic control through coordinated interaction of amygdala, theory of mind and extended pain matrix brain regions. *NeuroImage 114*, 105–119.

Buckner RL, Carroll DC. (2007). Self-projection and the brain. *Trends in Cognitive Sciences, 11*, 49–57.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 181–253.

Decety, J., & Sommerville, J. A. (2003). Shared representations between self and other: A social cognitive neuroscience view. *Trends in Cognitive Sciences, 7*, 527–533.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective Taking as Egocentric Anchoring and Adjustment. *Journal of Personality and Social Psychology, 87*(3), 327–339.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 360*(1456), 815–836.

Goldman, A. I. (1989). Interpretation psychologized. *Mind & Language, 4*, 161–185.

Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mind-reading*. Oxford: Oxford University Press.

Gordon, R. M. (1986). Folk psychology as simulation. *Mind & Language, 1*, 158–171. (Reprinted in *Mental simulation: Evaluations and applications*, by M. Davies, & T. Stone, Eds., 1995, Oxford: Blackwell. Reprinted in *Mind and cognition: An anthology*, 2nd ed., by W. Lycan, Ed., 1998, Blackwell Publishing).

Gordon, R. M. (1992). Reply to Stich and Nichols. *Mind & Language, 7*, 87. (Reprinted in *Folk psychology: The theory of mind debate*, by M. Davies, & T. Stone, Eds., 1995, Oxford: Blackwell).

Gordon, R. M. (1995). Simulation without introspection or inference from me to you. In M. Davies & T. Stone (Eds.), *Mental simulation: Evaluations and applications*. Oxford: Blackwell.

Gordon, R. M. (1996). Radical' simulationism. In P. Carruthers & P. Smith (Eds.), *Theories of theories of mind*. Cambridge: Cambridge University Press.

Gilead, M., Boccagno, C., Silverman, M., Hassin, R.R., Weber, J., & Ochsner, K.N., (2016). Self-regulation via neural simulation PNAS 113(36), 10037–10042.

Heal, J. (1986). Replication and functionalism. In J. Butterfield (Ed.), *Language, mind and logic* (pp. 135–150). Cambridge: Cambridge University Press.

Helmholtz, H. v. (1856). Treatise of physiological optics: Concerning the perceptions in general. In *Classics in psychology* (pp. 79–127). New York, NY: Philosophical Library.

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences, 29*, 105–110.

Jeannerod, M., & Pacherie, E. (2004). Agency, simulation and self-identification. *Mind & Language, 19*, 113–146.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron, 79*, 836–848.

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology, 67*, 596–610.

Perner, J., & Roessler, J. (2010). Teleology and causal reasoning in children's theory of mind. In J. Aguilar & A. Buckareff (Eds.), *Causing human action: New perspectives on the causal theory of action. A Bradford Book* (pp. 199–228). Cambridge, MA: The MIT Press.

Perry, J. (1979). The problem of the essential indexical. *Noûs, 13*, 3–12.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87.

Ross, L., Greene, D., and House, P. (1977). The false consensus effect: An egocentric bias in social perception and attribution processes. *Journal Experimental Social Psychology 13*, 279–301.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding*. Hillsdale, NJ: Erlbaum.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Tamir, D. I., Mitchell, J. P. (2010). Neural correlates of anchoring-and-adjustment during mentalizing. *Proceedings of the National Academy of Sciences of the United States of America, 107*(24), 10827–10832.

Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J., & Kirmayer, L. J. (2019). Thinking through other minds: A variational approach to cognition and culture. *The Behavioral and Brain Sciences, 30*, 1–97.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103–128.

Yildirim, I., Siegel, M., & Tenenbaum, J.B. (2020). The Cognitive Neurosciences, 6th edition, Gazzaniga, Mangun, Poeppel (Editors). MIT Press.

Mental Files and Teleology

Josef Perner, Markus Aichhorn, Matthias G. Tholen, and Matthias Schurz

Introduction

In this chapter, we make a plea for rethinking our scientific view of folk psychology and the role of mentalizing. We use arguments from philosophy of action, behavioral data from children's cognitive development, and brain imaging in adults. We see limited use of mental state ascription or mentalizing for understanding others and their actions. Mentalizing becomes critical when different perspectives have to be considered. Indeed, a central characteristic of the mind is to take a perspective on the world. As long as agents' perspectives do not differ, appeals to mentalizing are convenient but not necessary and often void of substance. We motivate and justify this view in section "Bounded Mentalism".

Although the notion of perspective is central to understanding the mind, its use in different fields looks promiscuous. For instance, how can we square our prototypic notion of perspective as visual perspective with linguists' claim that choosing a particular label for something puts a specific perspective on it (e.g., a creature as "the family dog" or "the destroyer of shoes," Clark, 1997). Mental files theory is a

J. Perner (✉)
Centre for Cognitive Neuroscience, University of Salzburg, Salzburg, Austria

Department of Psychology, University of Salzburg, Salzburg, Austria
e-mail: josef.perner@sbg.ac.at

M. Aichhorn · M. G. Tholen
Centre for Cognitive Neuroscience, University of Salzburg, Salzburg, Austria
e-mail: markus.aichhorn@sbg.ac.at; matthias.tholen@sbg.ac.at

M. Schurz
Wellcome Centre for Integrative Neuroimaging, Department of Experimental Psychology, University of Oxford, Oxford, UK

Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen, Nijmegen, The Netherlands

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_13

contemporary way to capture the common core of the interdisciplinary use of "perspective," and in the following, we adopt Recanati's (2012) terminology to explain what this common core is. A mental file is a representation of an entity, its referent. It has the function to track this referent over time and register facts (knowledge) about it. Typically, one has only one file for each object, but with our interest in perspectives, multiple filing, where more than one file is deployed for the same object, is most relevant. Multiple filing generates *coreferential files*, i.e., files that have the same referent.

Such cases frequently arise when one fails to recognize someone immediately. For instance (Recanati, 2012), one sees a man mow his lawn. When he turns, one identifies him as Noam Chomsky. So, one had briefly entertained two coreferential files of Chomsky capturing different perspectives. One file showed a stranger mowing his lawn, the other the famous linguist. To identify the stranger as Chomsky, one has to capture their identity by *linking* the two files.[1] Linking has to establish numerical identity (there is only one person out in the world) and accessibility of information (what holds true of the stranger, that he is mowing his lawn, also holds true of Chomsky and vice versa). Also calling something "the family dog" or "the destroyer of shoes" creates different files representing different perspectives of the same animal. To represent identity, the files have to be linked. A surprising and empirically testable consequence of this analysis is that understanding perspective is closely related to understanding identity. Different perspectives are different ways of presenting the same object or scene and, as Frege (1892) has made clear with his sense-reference distinction, identity statements inform that the two terms are different modes of presentation (senses) of a single object (referent).

How mental files can capture a person's false belief is illustrated in the classic false-belief story of *Mistaken Max* (Wimmer & Perner, 1983) as shown in Fig. 1: Max puts his toy car in box 1, fails to witness its transfer to box 2, and, therefore, falsely believes that his car is still in box 1. Two coreferential files can capture how Max conceives of his car differently from one's own conception. A *regular* file is used for one's own reasoning showing his car in box 2, while a coreferential *vicarious* file captures Max's belief by showing his car in box 1. This vicarious file needs to be indexed to Max to make clear whose perspective it carries. It also needs to be linked to the corresponding regular file of the toy to capture the fact that the toy, of which Max thinks it is still in box 1 (represented on the vicarious file), is the very same car, of which one knows that it is in box 2 (represented on one's regular file). Thus, the linking of vicarious files is not to enable free information flow as in the case of regular files but to ensure numerical identity. Nevertheless, the ability to

[1] The obvious alternative is to merge the two files into one containing all information from both files. One advantage of linking separate files lies in easier error correction. Should it turn out that, after all, the man wasn't Chomsky, then the link between stranger-file and Chomsky-file can simply be cut and each person remains associated with the appropriate information. Moreover, Anderson and Hastie (1974) have shown that people after receiving identity information tend to keep separate representations of the same individual before merging them. For simplicity's sake, we will therefore only talk about linking coreferential files.

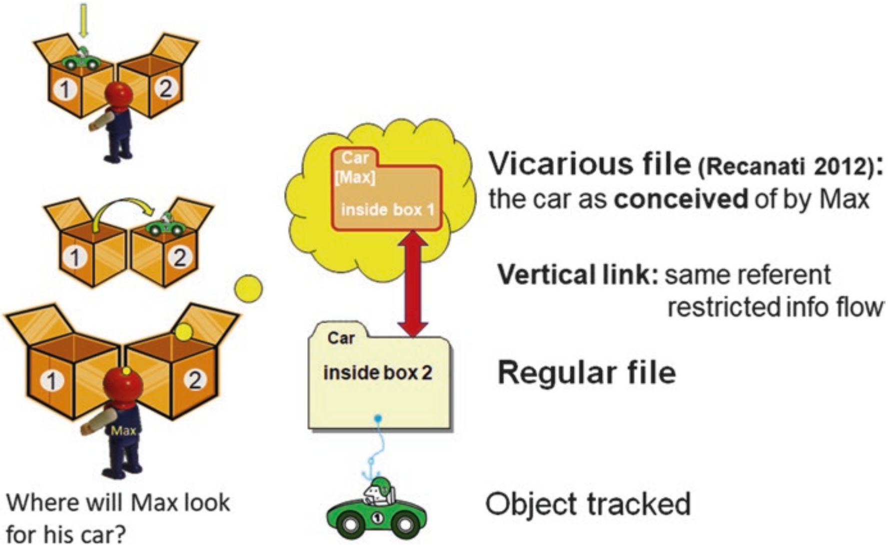


Fig. 1 The Mistaken Max false-belief scenario (free after Wimmer & Perner, 1983)

understand belief and the ability to understand identity statements both depend on the ability to link coreferential files. This can explain why identity and belief are understood by children in unison around the age of 4 years (Perner, Mauer, & Hildenbrand, 2011).

In section "Evidence", we give an overview of these findings woven in with the corresponding neurocognitive evidence that the exercise of these developmentally co-emerging abilities also activates a common brain region, the left inferior parietal lobe (IPL), and, less consistently, precuneus. The remarkable consistency of being acquired by children at the same age and activating a common cerebral region indicates a common cognitive basis.

Neurocognitive evidence shows that not all tasks that are currently classified as measuring "theory of mind" activate the same brain regions. In particular, processing false beliefs (FB) activates a region in the dorsal part of the left temporoparietal junction (TPJ) that lies within the inferior parietal lobe (IPL[2]), a region that is not typically activated by any other kind of theory of mind task (Schurz, Radua, Aichhorn, Richlan, & Perner, 2014). In the left hemisphere, this special region (left

[2] The terms "TPJ" and "IPL" are sometimes used inconsistently and require some clarification. As we have found in a literature review (Schurz, Tholen, Perner, Mars, & Sallet, 2017), researchers commonly use "TPJ" to refer to both structures of the Inferior Parietal (e.g., Angular Gyrus) and the Temporal Lobe (e.g., posterior Superior Temporal Sulcus/Middle Temporal Gyrus). The label "IPL," on the other hand, refers to an anatomical area found in standard brain atlases (i.e., gyral parcellations). The IPL is usually assumed to be confined by the Inferior Parietal Sulcus dorsally and the Lateral Sulcus ventrally. In the fMRI studies we review in this chapter, we used the label "IPL" as defined in the popular AAL atlas of the brain (Tzourio-Mazoyer et al., 2002).

IPL) overlaps with a region activated by several tasks in which identity plays a central role—all of which require linking of mental files: identity statements, arithmetic equations, person identification, etc. To note, corresponding tasks for children are all mastered at the same age as the FB task around 4 years, which affirms that linking of mental files is an important cognitive process. Having identified a specific brain region for linking of files helps explain why so many seemingly unrelated tasks are mastered at the same age. This coincidence would remain a mystery if these tasks were handled by independent domain-specific neural networks.

This developmental and neural evidence strongly suggests that we approach the interpretation of human action differently when no perspective differences are involved than when there are. This needs to be taken into account in our theories of social cognition.

Bounded Mentalism

We use *mentalism* for the rash tendency to defer to mental state attribution as *the prime* way of understanding people's (or other agents') conduct. It has become commonplace to equate social cognition with the use of a "theory of mind" (or belief-desire psychology), which activates a fairly uniform "mentalizing network" in the brain (Van Overwalle, 2009). However, given the fact that different parts of this network are used for different classes of tasks (Gobbini, Koralek, Bryan, Montgomery, & Haxby, 2007; Molenberghs, Johnson, Henry, & Mattingley, 2016; Schaafsma, Pfaff, Spunt, & Adolphs, 2015; Schurz et al., 2014), we might want to question this widespread assumption. To understand these differences, it might help differentiate the different philosophical approaches. One theory claims that we are using a theory of mind ("theory" theory) by attributing nonobservable mental states for predicting and explaining behavior (Premack & Woodruff, 1978). This theory consists of an explicit grasp of the lawful regularities of agents' minds and actions (Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1992). It has been opposed by simulation theory claiming we imagine ourselves being in other people's situation and observe how we react to it (Gallese & Goldman, 1998; Gordon, 1986; Heal, 1986). Although we may, on occasion, use such a technique, we have no awareness that we do use it on a regular basis.

A third approach is inspired by the philosophy of action (Alvarez, 2018; Anscombe, 1957; Davidson, 1963; Raz, 1999), where actions are defined by reasons. One acts intentionally when one acts for a reason that could be articulated in answer to the question "why did you do that?" A reason is any fact that speaks in favor of the action (Scanlon, 1998). Perner and Roessler (2010; Roessler & Perner, 2013) proposed this as the basis of our folk psychology of action. They called it *teleology* because intentional action is aimed at improving conditions (the telos or goal) and use practical reasoning to figure out how to achieve this. Our everyday conversations illustrate this nicely: YOU: "What did you do on the weekend?"—ME: "I decided to hike up the big mountain."—"Why?"—"One gets the best view from there."—"But isn't it too far to walk there?"—"I took the bus to the parking lot below the peak."

There are several key features to note about this brief exchange:

1. Reasons are objective facts: Being teleologists, we made sense of my weekend exploits in terms of justifying reasons provided by "worldly" facts (not by mental states): There is a worthwhile (good) thing to be had: the best view of the region; which provides a good reason to go there. This is treated between us as an objective fact and not a subjective, idiosyncratic desire. And so is the fact that using the bus and hiking up to the top is an objectively good way of getting to enjoy that view.
2. The goal is the good—Aristotle (Charles, 2012): I made my actions intelligible to you because they were designed to achieve something good, i.e., a more enjoyable view than if I had stayed at home. In fact, had I said "cause I *wanted* to go there" (a perfectly informative answer for a theory user), you would have not found my answer illuminating. You probably would have found me stroppy and unwilling to continue the conversation. My answer would have been disappointing because it would have failed to make clear what good was supposed to come of my action. I had a goal, okay, the mountain top but nothing attractive about it.
3. Choice: Our conversation shares the tacit assumption that I chose to hike up the mountain (Raz, 1999). I was not drawn to do so by an uncontrollable desire following some law of nature.
4. Shared facts: Since reasons for action are objective, all participants can be made aware of them. Thus, there is no need to take anyone's subjective perspective to understand what she is doing. Everybody can look at the same facts from their respective first-person perspective and see what needs doing. Thus, the starting point for action is not a private desire but a shared view of how things can be improved, a goal. This makes teleologists poised for cooperation (Perner & Esken, 2015). In other words, teleologists look what needs doing and assume that someone, who is aware of the facts, will do what is needed. Who that will be depends on additional facts: who is the most competent to do it; who is responsible for it; etc.

Thus, teleology differs radically from use of a theory as advocated by Gopnik and Meltzoff (1997, p. 126):

> These tenets [the main tenets of the 'adult theory'] are perhaps best summarized by the "practical syllogism": "If a psychological agent wants event *y* and believes that action *x* will cause event *y*, he will do *x*." Many philosophers have argued that the practical syllogism is the basic explanatory schema of folk psychology.

More generally speaking, a user of a theory of mind infers an agent's mental states, in particular beliefs and desires, from what can be observed about the agent. Perusing the practical syllogism, the agent's likely behavior can be predicted. There is no mention that agents act on the basis of *reasons* to achieve *worthwhile* goals, or choose their actions. Perhaps, reasons, values, and free choice are misplaced in a scientific psychology, but they surely are at the core of our folk psychology. Teleology gives room for these aspects of our folk psychology, which have been neglected by the belief-desire psychology of theory of mind.

Teleology differs, to its disadvantage, in another important way from theory of mind. Teleology has no inbuilt way of coping with unresolved differences of perspective. The well-known false-belief test for children with Mistaken Max (see above) makes that clear. A teleologist would consider the problem posed: Max needs to get to his toy, and the best way for him to achieve this is to move to where the toy currently is. The teleologist concludes that Max being sensible and rational will go to that location. This is the wrong answer, easily avoided by using a theory of mind. Max doesn't know that his chocolate has been moved and mistakenly thinks it is still where he had put it. He will, therefore, go to where he thinks it is, not where he should go to find it.

Clearly, teleology needs some additional provisions to be able to provide the adult answers in such error cases. This can be achieved by applying it *counterfactually* to the informational state of the agent ("teleology-in-perspective": Perner & Roessler, 2010; Rafetseder, O'Brian, Leahy, & Perner, 2018). That is, one asks what one would have reason to do, if the world were as the agent's information presents it. For instance, Max has witnessed putting his toy into box 1, going out to play, returning. If this *were* all that had happened, then the toy *would still be* in box 1 and Max *would have* good reasons to look for it there.

Ironically, teleology's initial weakness of needing an additional fix to cover different perspectives may work in its favor explaining developmental data. In fact, 3-year olds give the teleologist's wrong answer to the false-belief problem with confidence (Ruffman, Garnham, Import, & Connolly, 2001). Only around 4 years (Wellman, Cross, & Watson, 2001) do they make the additional step and apply teleology within Max's perspective and predict correctly that he will go where according to his perspective he should go.

From these considerations, two developmental consequences follow. If children grow up as teleologists, there will be a sharp distinction in their mastering social tasks between those tasks that involve perspective differences (perspective tasks) and those that do not. The other consequence is that mastery of perspective tasks should go hand in hand with counterfactual reasoning ability.

Neither of these consequences follows naturally from the use of a theory of mind understood as a belief-desire psychology. Attribution of a mental state is similar, whether the other agent's perspective does or does not differ from one's own. For instance, when using a theory of mind for Mistaken Max, one has to apply attributional principles like "seeing leads to believing." So Max seeing his toy being put into box 1 leads to "Max believes his toy is in box 1." Since Max does not witness its move to box 2, his belief that it is in box 1 remains unchanged, and he is stuck with a false belief that differs from one's own view. If in a true belief condition, Max sees the toy being transferred, which changes his belief from "box 1" to "box 2." Hence, application of the theory does not look very different in the two cases: so why should children find one easier than the other task? The obvious difference, of course, is the fact that only in the false-belief case different perspectives are involved, which makes this task so much more difficult. But how the belief content relates to reality is not taken into account by a theory of mind analysis. Hence, in order to explain the massive difference in difficulty between true- and false-belief tasks, researchers had to look for theory-extraneous factors. In particular, interference between the toy's real and believed location is thought to exceed the younger children's executive strength to inhibit this interference (Baillargeon, Scott, & He, 2010; Setoh, Scott, & Baillargeon, 2016).

Our analysis also shows that theory of mind has, even when a belief turns out false, no obvious place for counterfactual conditionals as required for teleology-in-perspective. For instance, since Max did not witness the transfer of his toy, then the transfer did not happen from his perspective. The teleologist, therefore, has to reason counterfactually, "if the transfer had not happened, Max's toy would still be in its original location and Max would have a good reason to go there to get it." Theory of mind has no use of such reasoning. Since Max failed to witness the transfer, the attributional principle "seeing leads to believing" does not apply and so his belief, that the toy is in box 1, will not be updated. Since the principle is applied to Max's actual lack of information, no counterfactual assumptions have to be made.

To take stock, we urged to rethink the widespread acceptance that people explain and predict each other's behavior with a theory of mind. Such a theory allows one to use lawful generalizations to infer from an agent's observed behavior and circumstances the agent's mental states. From knowledge of these mental states, one can predict the agent's future behavior. But this is not the way we explain our own actions, as our mountain view example above showed. We explain why the goal of our action was attractive and why the action was a good way to achieve the goal: pure teleology. Pure teleology needs to be extended by being used counterfactually to account for differences of perspective. It, therefore, implies a strict separation of cases of a shared world and cases where agents entertain different perspectives. This implication is borne out by the fact that children master all kinds of different perspective tasks at about the same age. Moreover, this developmental synchrony extends beyond social perspective taking and is also reflected in understanding identity or false direction signs. To explain this connection, we use mental files theory and conclude that the common factor is the ability to link coreferential files. This accounts for the common developmental trajectory of perspective and identity tasks and why they activate a common brain area in adults. We now present these data in more detail.

Evidence

We can classify tasks according to our two examples above. Identifying a person involves establishing the identity of the referent of two coreferential files, e.g., the stranger mowing his lawn with Noam Chomsky. Let us call these *identity tasks*. Tracking another person's perspective requires intentionally deploying two files for the same entity, e.g., for Max's toy. We refer to them as *perspective tasks*. In both cases, *linking* of coreferential files is essential. In the following, we introduce different tasks that have both been used with children and in brain imaging with adults. They show a remarkable consistency of being acquired by children at the same age and of activating a common cerebral region in the left IPL. This strongly indicates a common cognitive basis.

Perspective Tasks

Visual Perspective

Developmental work on visual perspective taking introduced an important distinction of levels (Flavell, Everett, Croft, & Flavell, 1981; Masangkay et al., 1974). A Level 1 task requires children to distinguish what they can see from what others can see, and they pass this task at the age of 2 years (Moll & Tomasello, 2006). Level 1 is not a perspective task in our sense since they do not require coreferential filing. As shown on the left side of Fig. 2, a relevant difference can be encoded on each person's file by stating which object is or is not in each person's visual field. The sitting avatar can see the two large blocks but not the tiny one hidden from him by one of the big blocks. There is no need to deploy a vicarious block-file indexed to the avatar because the information that the block is outside his visual field can be

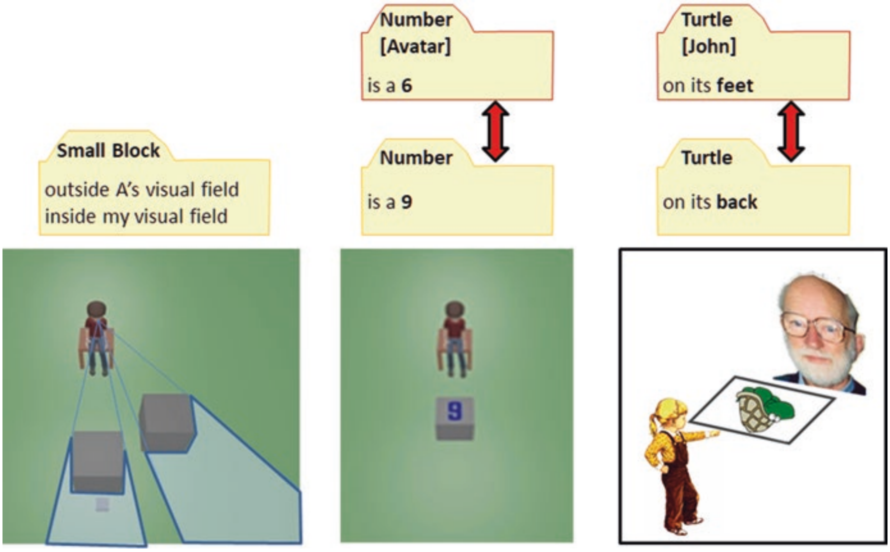


Fig. 2 A Level 1 (left) and a Level 2 (center) visual perspective-taking task used in brain imaging studies, and one Level 2 task used by John Flavell for children (right). The corresponding mental file analyses are shown above each picture. The bracketed names on vicarious files show to whom the file is indexed. The visual lines and areas of occlusion in the left picture were not part of the original experimental stimuli

stated on one's regular file. The figure makes clear that one can see that the avatar cannot see the block, just like a lamp in his place could not illuminate the little block (see Schurz et al., 2015). We would not think of capturing this fact with a vicarious file indexed to the lamp.

Level 2 tasks require the understanding that different people looking at the same objects may see them differently; e.g., in the right picture of Fig. 2, child and experimenter are both looking at the same drawing of a turtle. Yet, one of them will see it as standing on its feet and the other as lying on its back. Hence, two files need to be deployed for the turtle to encode the otherwise incompatible propositions "it is on its feet" and "it is on its back." This task is mastered around 4 years, and performance correlates with the false-belief task (Hamilton, Brindley, & Frith, 2009). The central panel in the figure shows a Level 2 problem for adults (Surtees, Apperly, & Samson, 2013) used in brain imaging studies (Schurz, 2015). Again, the perspectival facts that the number on the block is a "6" (for the avatar) and a "9" (for us) cannot be both registered on a regular file without contradiction. So a vicarious file of the number is needed for the avatar.

Brain imaging studies of visual perspective taking (vPT) used mostly Level 2 tasks. Visual perspective has been the prototype for investigating understanding of others' perceptions (Piaget & Inhelder, 1948/1956). Perceptions are mental states. Hence, it is surprising that vPT has not featured in representative meta-analyses of theory of mind or mentalizing (Molenberghs et al., 2016; Schurz et al., 2014). Even more surprising, when Schurz, Aichhorn, Martin, and Perner (2013) tested for meta-analytic overlap between vPT and FB tasks, there was hardly any overlap (see Fig. 3).[3] Nor was there much overlap with other theory of mind tasks (Arora, Schurz, & Perner, 2017).

As Fig. 3 shows, the only overlap was in precuneus, left IPL, and in left middle occipital gyrus. This gave rise to the idea that these areas might be specifically responsible for dealing with perspective differences, one aspect shared by false beliefs and visual perspectives. This view will be corroborated in the following for left IPL and, within limits, for precuneus.

False Beliefs, Direction Signs, and Photos

Children's performance on the typical false-belief stories shows a stable developmental transition between 3 and 5 years (Wellman et al., 2001) from mostly incorrect answers (Max will look for his toy where it actually is) to mostly correct

[3] Of the 14 studies in this analysis, three used Level 1 and all others were clear cases of Level 2 perspective taking. We checked whether the two groups tended to activate particular regions differently, but there was no noticeable difference. Since children pass Level 1 tasks earlier, presumably because they can give correct answers without an understanding of perspective, one would have expected a difference. However, adults might spontaneously concern themselves with the appearance of what the other person sees, which would activate perspective processing areas just as Level 2 tasks would.

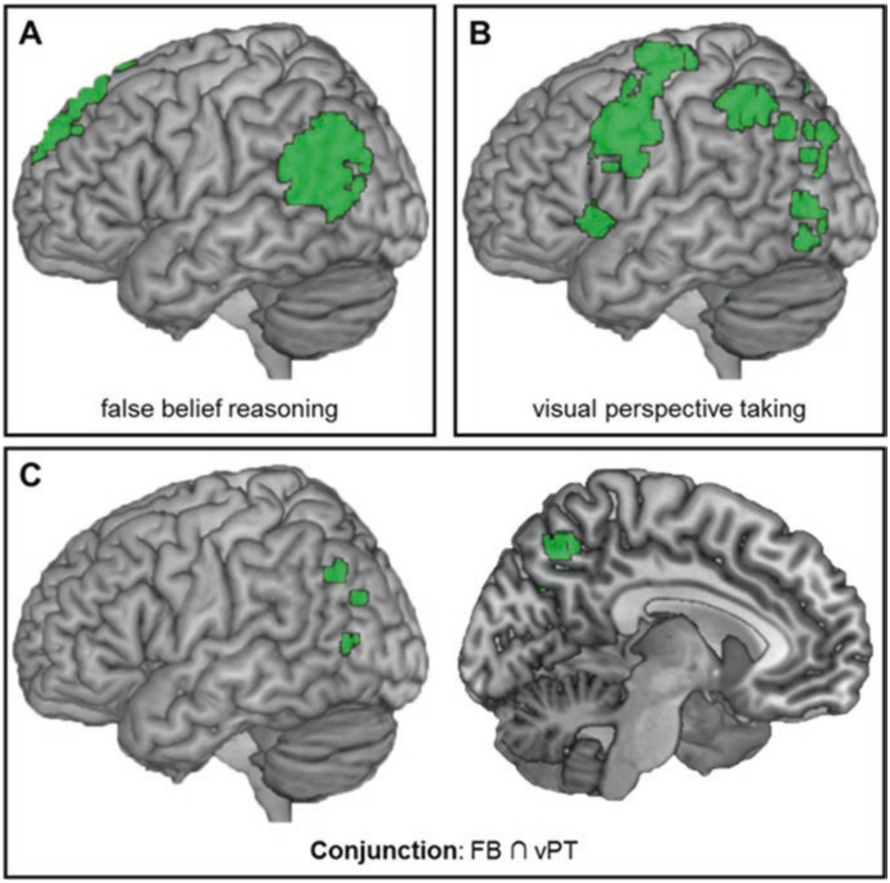


Fig. 3 Results of meta-analyses for false-belief reasoning (**a**) and visual perspective taking (**b**) (taken from Schurz et al., 2013). Results of a conjunction analysis searching for brain areas active for false-belief reasoning AND visual perspective taking (**c**). All maps were thresholded at a voxel-wise threshold of $p < 0.005$ uncorrected and a cluster extent threshold of 10 voxels

answers (he will look where he thinks it is). However, there has been a flourishing series of publications of ever earlier evidence for belief attribution using looking behavior as an indirect test. Clements and Perner (1994) found a dissociation between correct visual anticipation of the agent's mistaken action and wrong prediction when asked the standard test question. Despite their correct looking, children were convinced of the correctness of their wrong verbal response (Ruffman et al., 2001). Correct visual anticipation could be shown just before children's third birthday but not earlier. All of seven known replication reports confirmed these findings (Kulke & Rakoczy, 2018). Much earlier evidence came from similar studies that avoided any verbal interaction. Children had to infer what the agent wanted from repeated successful retrieval of the target object. Using looking time to measure violation of expectation yielded evidence around 14 months (Onishi &

Baillargeon, 2005; Surian, Caldi, & Sperber, 2007, and many others). Correct anticipatory looking was found at 2 years (Southgate et al., 2007) or 18 months (Neumann, Thoermer, & Sodian, 2008; Thoermer, Sodian, Vuori, Perst, & Kristen, 2012) provided the target object was removed from the scene (disappear condition) to eliminate any distraction of the object's real location (and—presumably—application of basic teleology). The replicability of this early evidence has, though, come under sustained criticism. Across several different study paradigms, only about half the studies found the effects (Kulke & Rakoczy, 2018).

The nature of findings and the volatility of replicability fit a version of mental files theory ("unlinked vicarious files": Perner, 2016). Before passing the standard false-belief test, children are able to form vicarious files but cannot link them. This gives infants information about another agent's perspective in the vicarious file. So, when they happen to track the object with that file, its information takes hold of their practical reasoning. However, as the file is not linked to their regular file, they cannot use it when tracking the object with their regular file. This leads to haphazard use of the vicarious file sufficient for above chance performance but not intentional use of an alternative to their own perspective. We conclude that consistent performance on the traditional verbal false-belief test is, after all, a good indicator of children's ability to link coreferential files.

Most studies using the traditional test used the unexpected transfer paradigm of the original Mistaken Max story (Wimmer & Perner, 1983). Zaitchik (1990) designed a nonmental version of this task dubbed the *"false"-photo* task. Its structure was highly similar: a target object is in location 1, and a Polaroid photo is taken, the object changes to location 2, and children are asked: "In the picture, where is the object?" As a group, 3- to 5-year olds found the question as difficult as the FB question, indicating problems with representation in general and not just false belief. A later review of 14 studies (Perner & Leekam, 2008) confirmed the same difficulty but showed a surprising absence of significant correlations between the FB and the "false-photo" task, especially when controlled for age or verbal intelligence. This finding should not be surprising if these tasks rely on different cognitive functions despite their surface similarity.

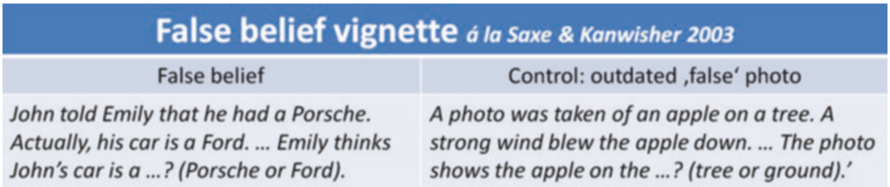
The strong surface similarity between the false-belief and false-photo task covered up an important difference. The photo is not false; it is a photo of a bygone time; or else our holiday photos would all be false! Unfortunately, we are not aware of any mental files analysis of photos or pictures. This denies us sound theoretical grounds to claim that photos do not need linking of vicarious with regular files. Nevertheless, there are suggestive differences between photos and beliefs. Beliefs, as Ramsey (1931; Jackson & Braddon-Mitchell, 1998) once put it, are *maps by which we steer*. Our beliefs show us which actions to take to attain our goals. This is not required of the photo in the "false"-photo task. One is asked to describe events in the photo independent of what is the case with the same objects in reality. For beliefs, the relation between what they show about reality and reality is essential. And vicarious files are exquisitely apt to capture the function of beliefs. Their vicarious use makes one steer according to another person's belief. That is why we need vicarious files for representing belief. They are not needed for describing the content of photos. Hence, the photo task does not require vicarious files, and consequently no linking them with regular files.

Direction signs are a different matter. Like photos, they are nonmental representations, but like beliefs, they represent to steer. For instance, when the sign for the ice cream van points to behind the church (Parkin, 1994), it helps to find the van there. Also, the question, "Where does the sign show that the van is?" cannot be answered by looking for the van in the sign and one has to interpret the sign in relation to its environment, whereas the corresponding question about the photo can be answered by just looking for the van in the photo. To understand what the sign shows one needs to understand how it affects one's actions and therefore a vicarious file for the van is asked for. And it needs to be linked with the corresponding regular file. In line with this analysis, data from eight false-sign studies (review by Perner & Leekam, 2008, Table 2) showed performance on false sign correlated strongly with performance on false-belief tasks independent of age or other covariates, and even with performance on the "false"-photo task controlled (Leekam, Perner, Healey, & Sewell, 2008). The very same pattern of results also holds for children with autism spectrum disorders. Performance on false-belief and "false"-photo tasks do not go hand in hand (Leekam & Perner, 1991), whereas children perform comparably on false-belief and false-sign tasks (Iao & Leekam, 2014).

The majority of neurocognitive studies on false-belief ascription used Saxe and Kanwisher's (2003) contrast between false-belief and "false"-photo vignettes (see Table 1). The meta-analysis by Schurz et al. (2014) showed marked activation differences for six kinds of theory of mind research. Two areas were activated only by FB-vignettes, the dorsal part of the TPJ within the IPL in both hemispheres. On the left side, the test point was at "−46,−63,41," a mere 6.5 mm distant from the peak voxel of the overlap between FB and vPT (Schurz et al., 2013; see row 12a in Table 3 below). This confirms the idea that this area specializes in perspective computation, as only FB but no other theory of mind test activated this region.

To firm up this conjecture further, we tested whether false-sign vignettes, which for children are as difficult as false-belief stories, also activate this area. Perner, Aichhorn, Kronbichler, Staffen, and Ladurner (2006) had four tightly controlled conditions with, e.g., an object in location X but depending on condition *believed* to be in location Y (FB), *indicated* to be in Y (false sign), *shown* in Y (Photo), and having *been* in Y (temporal change). The signal plots in Fig. 4 show that FB vignettes activated right TPJ significantly more than any of the other conditions, supporting

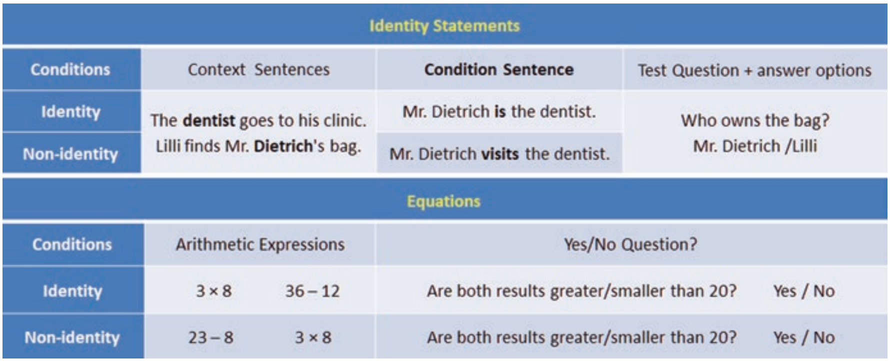
Table 1 A shortened version of a false-belief and a "false"-photo vignette used by Saxe and Kanwisher (2003



False belief vignette *á la Saxe & Kanwisher 2003*

False belief	Control: outdated ‚false' photo
John told Emily that he had a Porsche. Actually, his car is a Ford. ... Emily thinks John's car is a ...? (Porsche or Ford).	*A photo was taken of an apple on a tree. A strong wind blew the apple down. ... The photo shows the apple on the ...? (tree or ground).'*

Table 2 Procedure for identity statements and equations

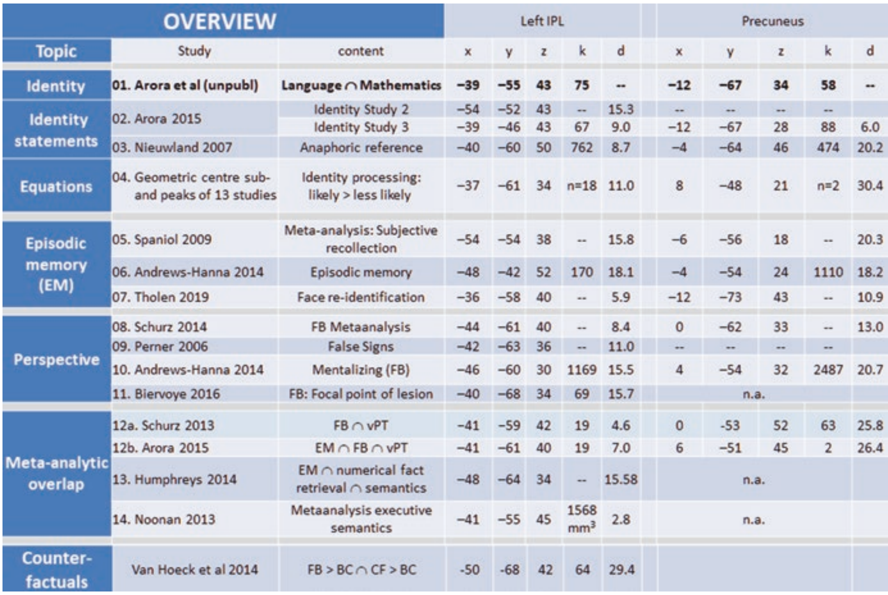


Identity Statements			
Conditions	Context Sentences	**Condition Sentence**	Test Question + answer options
Identity	The **dentist** goes to his clinic. Lilli finds Mr. **Dietrich**'s bag.	Mr. Dietrich **is** the dentist.	Who owns the bag? Mr. Dietrich /Lilli
Non-identity		Mr. Dietrich **visits** the dentist.	

Equations				
Conditions	Arithmetic Expressions		Yes/No Question?	
Identity	3 × 8	36 – 12	Are both results greater/smaller than 20?	Yes / No
Non-identity	23 – 8	3 × 8	Are both results greater/smaller than 20?	Yes / No

The two arithmetic formulae yield the same number in the identity condition (e.g., **24**) but not in the nonidentity condition **(15, 24)**

Table 3 MNI coordinates of peaks or subpeaks closest to the identity conjunction (identity statements ∩ equations) in left IPL and precuneus of all relevant studies



OVERVIEW			Left IPL					Precuneus				
Topic	Study	content	x	y	z	k	d	x	y	z	k	d
Identity	**01. Arora et al (unpubl)**	**Language ∩ Mathematics**	**−39**	**−55**	**43**	**75**	**--**	**−12**	**−67**	**34**	**58**	**--**
Identity statements	02. Arora 2015	Identity Study 2	−54	−52	43	--	15.3	--	--	--	--	
		Identity Study 3	−39	−46	43	67	9.0	−12	−67	28	88	6.0
	03. Nieuwland 2007	Anaphoric reference	−40	−60	50	762	8.7	−4	−64	46	474	20.2
Equations	04. Geometric centre sub- and peaks of 13 studies	Identity processing: likely > less likely	−37	−61	34	n=18	11.0	8	−48	21	n=2	30.4
Episodic memory (EM)	05. Spaniol 2009	Meta-analysis: Subjective recollection	−54	−54	38	--	15.8	−6	−56	18	--	20.3
	06. Andrews-Hanna 2014	Episodic memory	−48	−42	52	170	18.1	−4	−54	24	1110	18.2
	07. Tholen 2019	Face re-identification	−36	−58	40	--	5.9	−12	−73	43	--	10.9
Perspective	08. Schurz 2014	FB Metaanalysis	−44	−61	40	--	8.4	0	−62	33	--	13.0
	09. Perner 2006	False Signs	−42	−63	36	--	11.0	--	--	--	--	
	10. Andrews-Hanna 2014	Mentalizing (FB)	−46	−60	30	1169	15.5	4	−54	32	2487	20.7
	11. Biervoye 2016	FB: Focal point of lesion	−40	−68	34	69	15.7	n.a.				
Meta-analytic overlap	12a. Schurz 2013	FB ∩ vPT	−41	−59	42	19	4.6	0	-53	52	63	25.8
	12b. Arora 2015	EM ∩ FB ∩ vPT	−41	−61	40	19	7.0	6	−51	45	2	26.4
	13. Humphreys 2014	EM ∩ numerical fact retrieval ∩ semantics	−48	−64	34	--	15.58	n.a.				
	14. Noonan 2013	Metaanalysis executive semantics	−41	−55	45	1568 mm^3	2.8	n.a.				
Counter-factuals	Van Hoeck et al 2014	FB > BC ∩ CF > BC	-50	-68	42	64	29.4					

k **number of voxels,** *d* **difference in mm of study peak or subpeak to peak of identity conjunction**

Saxe and Kanwisher's (2003) claim that right TPJ is specifically associated with processing mental states like belief and thinking. On the left side, however, FB as well as the false sign activated the TPJ more strongly than the photo or the temporal change condition. This asymmetry was confirmed in a region of interest approach by Aichhorn et al. (2009). Moreover, the peak voxel of this area (row 09 in Table 3)

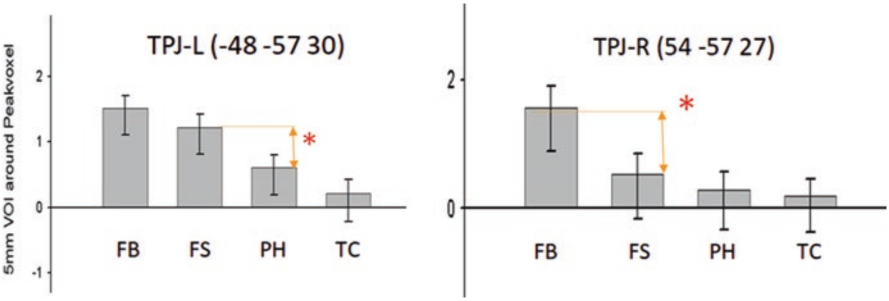


Fig. 4 Signal change and SE of the four conditions in spheres of 5 mm around peak voxels for FB > PH contrast. FB, false belief; FS, false sign; PH, "false" photo; TC, temporal change

was only 7.3 mm away from the peak voxel of the overlap between FB and vPT (row 12a in Table 3). This reconfirms our conjecture of this region's involvement in processing perspective differences.

The mental files analysis given earlier makes an even more ambitious, wider claim about this dorsal part of the TPJ in the IPL. It should not only be involved in perspective but also in identity computations. So we tested this claim in three different areas: linguistic identity statements, mathematical equations, and face recognition.

Identity Tasks

Verbal and Numerical Identity

In their "key experiment," Perner et al. (2011) had children mark a key with a yellow sticker for opening the yellow box and another key with a green sticker for the green box. As it turned out, this was the same key that opened both boxes, with a green on one and a yellow sticker on its other side. After discovering this identity, the children who failed the false-belief test denied that the key with the green marker visible would open the yellow box. The developmental trend is displayed in Fig. 5, left panel. Performance on the identity and the false-belief task also correlated strongly ($r = 0.57$) independently of age and performance on the dual function control task ($r_p = 0.36$), in which a key opening both boxes, but whose identity was never in doubt, was given the respective color markings on each of its sides.

In a "lost-and-found story," children saw a man described as *the firefighter* walk into a house. Then, a bag belonging to Mr. Mueller was found. After being told "Mr. Mueller **is** the firefighter," children, who failed the false-belief test, did not realize that the bag belongs to the firefighter. Performance on belief and identity tasks described a similar developmental trajectory (shown in Fig. 5, right panel) and

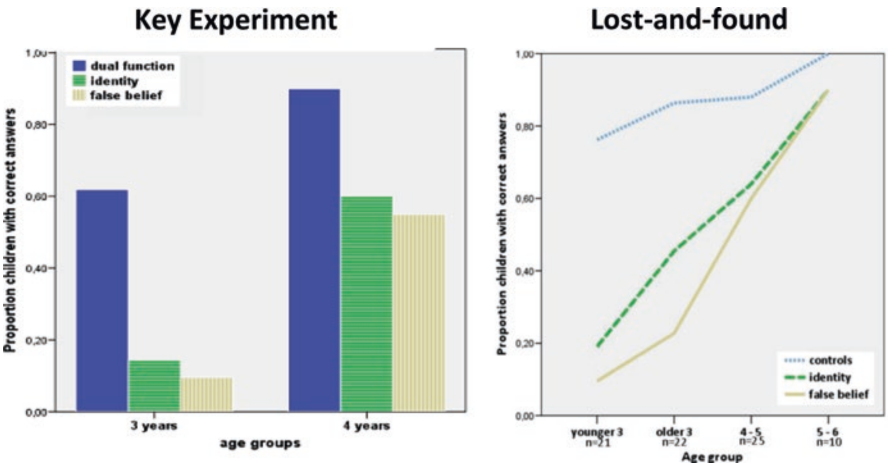


Fig. 5 The proportion of children passing two identity experiments (from Perner et al., 2011)

correlated strongly ($r = 0.68$) with age and verbal IQ controlled ($r_p = 0.50$).[4] These children's problems were specific to processing the identity information, because in one control task most could remember who the firefighter was, and when told "Mr. Mueller is a firefighter," in a second control task they understood that the bag belongs to the person dressed as a firefighter.

For our brain imaging studies, the leading question was whether this co-development would also be reflected in common neural structures specialized for linking coreferential files? We expected activation in left IPL and precuneus as these regions were activated by both, false-belief vignettes and visual perspective taking. To this end, we designed short identity vignettes and nonidentity control vignettes like the ones in Table 2. In both conditions, a person (the dentist) is introduced, and then an object of an unknown person (Mr. Dietrich) is found. In the identity condition, the owner of the object is identical to the first person (Mr. Dietrich is the dentist), while in the control condition, it is made clear that they are different people (Mr. Dietrich visits the dentist).

We also wanted to see whether the identity vs. nonidentity contrast for linguistic statements overlaps with the realization that two arithmetic formulae yield the same number.[5] According to Frege, the "=" sign of mathematical equations has the same semantics as the "is" in identity statements. To this end, participants were instructed to solve each of the two formulae presented on each trial to be able to answer occasional questions (lower panel of Table 2). The critical difference between conditions

[4] We have several unpublished student projects with many different variations of the identity problem. They all showed the same age trend and consistently correlations with the false-belief test.

[5] There are no developmental studies that show a correlation of comparable mathematical prowess with false-belief understanding. However, several studies (see Carey, 2009) show that children start to understand the cardinality principle of counting sets around 4 years, and Sarnecka and Wright (2013) found that with that principle children also understand equinumerosity.

was that the two formulae yielded the same number in the identity condition but different numbers in the control condition.

The imaging results are shown in Fig. 6. There were no lateral activations in the right hemisphere. Two of the activated areas correspond to areas of overlap in the false-belief and visual perspective-taking meta-analysis (Schurz et al., 2013): left IPL and precuneus. Their peak voxel coordinates are shown in row 01 of Table 3.

Person Identity

We have introduced the importance of linking coreferential files with the example of a stranger mowing his lawn. Upon identifying *that stranger* as *Noam Chomsky*, one has to link the newly deployed file for the stranger with one's long-existing Chomsky file. If the person had been a stranger, only a new file would have been deployed without linking it to another file. We captured this difference between identifying someone as a familiar person and having encountered a new person in our next study.

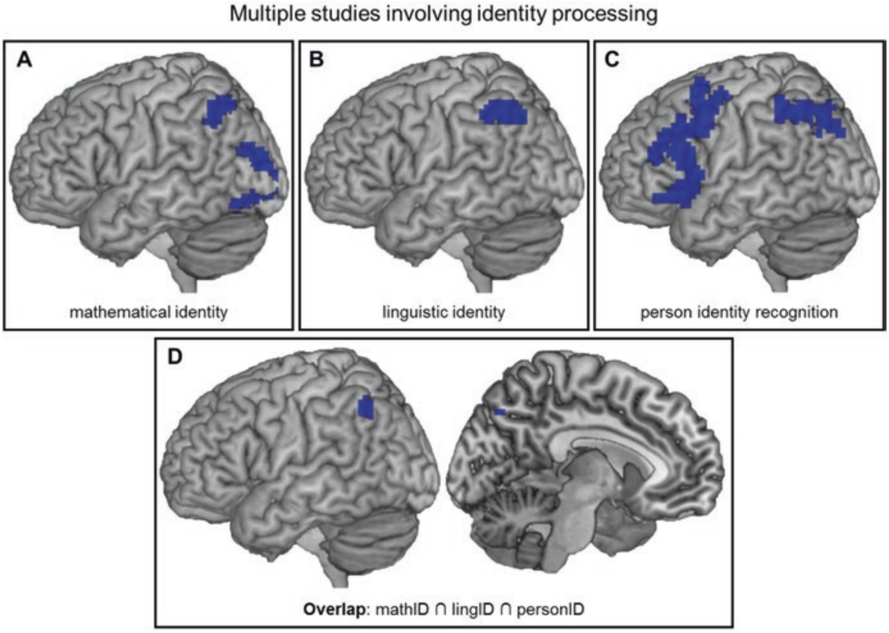

Multiple studies involving identity processing
A
mathematical identity
B
linguistic identity
C
person identity recognition
D
Overlap: mathID ∩ lingID ∩ personID

Fig. 6 Results of studies involving identity processing: mathematical identity (**a**) and linguistic identity (**b**) (taken from Arora et al. under revision) and person identity recognition (**c**) (taken from Tholen, Schurz, & Perner, 2019). Results of an overlap analysis searching for brain areas active for mathematical identity (mathID), linguistic identity (lingID), and person identity recognition (personID) (**d**). All maps were thresholded at a voxel-wise threshold of $p < 0.001$ uncorrected together with an FWE-corrected cluster threshold of $p < 0.05$

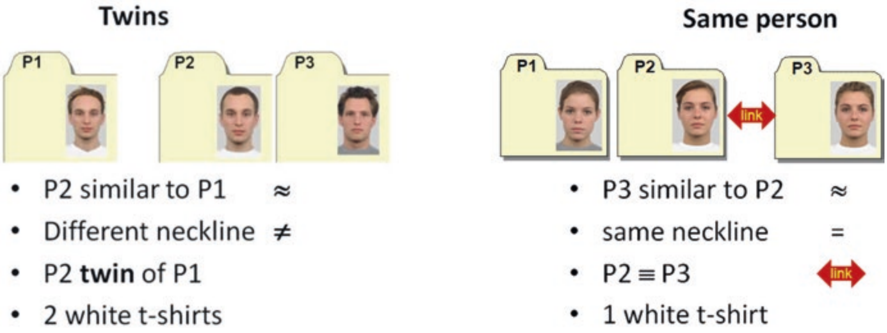


Fig. 7 The mental files analysis of two critical trials in the study by Tholen et al. (2019)

Trials consisted of a sequence of three passport portraits. On some trials, two of the three persons looked very similar. Subjects were instructed to check then whether the necklines of these persons were the same or different. If the same, they were the same person, and if different, they were similar looking twins. To ensure subjects attended to the identity of persons, they were asked on each trial how many of the people had a white T-shirt. Figure 7 depicts the mental files analysis of two critical trials. On the left side, Persons P1 and P2 look similar but have different necklines, so the two files should not be linked since they refer to different individuals. Since each of them has a white T-shirt but not the third person, the number of T-shirts is 2. On the trial shown on the right side, P2 and P3 look similar and have the same neckline, which means that these are two photos of the same person. Both photos show her but not the first person with a white T-shirt. The T-shirt count is thus 1.

Stronger BOLD activity was found for the same person than for the twins in medial prefrontal cortex, the left inferior frontal gyrus, thalamus, as well as bilateral IPL/TPJ, left precuneus, and the lateral occipital cortices. No brain regions showed a greater response for the reverse contrast. The predicted activation differences in left IPL and precuneus are shown in Fig. 6c. Both are within or contiguous with the conjunction of identity statements and equations. The closeness to the statement-equation conjunct is also shown in Table 3 (row 07); the peak voxels are within 5.9 mm for the left IPL and within 10.9 mm for the precuneus. Figure 6d shows the conjunct of all three identity tasks, statements, equations, and person identification in the left IPL and in precuneus.

Figure 8 also shows that this identity conjunct touches the overlap of false belief with visual perspective taking (green). The peak voxels of these overlap areas are shown in row 12a of Table 3. They are very close to the conjunction peaks in row 01, within 4.6 mm in the IPL and 25.8 mm in the precuneus.

Our findings have wider significance. For one, they identify the neural processes underlying new views (Darby & Caplan, 2016; Wilkinson, 2016) of misidentification syndromes, notably Capgras, which we would characterize as an inability to link files. Patients with Capgras delusion fail to identify certain individuals, even

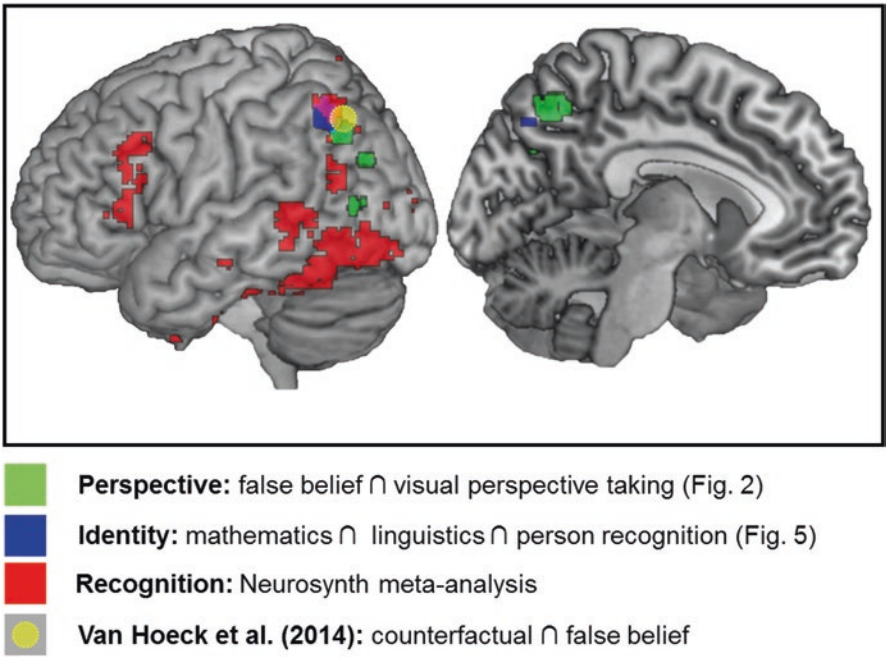

🟩 Perspective: false belief ∩ visual perspective taking (Fig. 2)
🟦 Identity: mathematics ∩ linguistics ∩ person recognition (Fig. 5)
🟥 Recognition: Neurosynth meta-analysis
🟡 Van Hoeck et al. (2014): counterfactual ∩ false belief

Fig. 8 Results of the overlap analysis between false-belief reasoning and visual perspective taking (green), the overlap analysis between mathematical identity, linguistic identity and person identity recognition (blue), neurosynth meta-analysis on "recognition" (red), and the peak voxel with 5 mm sphere (−36,−66,44) from the conjunction analysis between counterfactual reasoning and false-belief reasoning (taken from Van Hoeck et al., 2014) (yellow)

close family members (e.g., their wife), and despite an unimpaired ability to recognize the strong similarity (Hirstein & Ramachandran, 1997). As a consequence, patients conclude that the person, who resembles their wife so convincingly, must be an impostor (Ellis & Lewis, 2001).

These symptoms would follow from a problem with recognizing a person immediately together with an inability to link files for identification (Wilkinson, 2016). The similarity to the family member can be seen but not their identity. A file linking deficiency can also explain reduplication of impostors, where the impostor has again been replaced by another impostor and so forth (Capgras & Reboul-Lachaux, 1923; Wilkinson, 2016). Our emphasis on how perspective taking hinges on representing identity provides a direct explanation for why misidentification syndromes are associated with mentalizing deficiencies as noted by Gobbini and Haxby (2007) and Hirstein (2010: "misidentification syndromes as mindreading disorders").

Our findings also lead to relevant predictions for recognition memory. The Neurosynth map for "recognition" in Fig. 8 shows activations (red) primarily due to the contrast between recognized old items and new items. One of these areas falls squarely into the conjunction of identity sentences, equations, and face identification data (blue). The prediction is that this activation is due to items that were not immediately recognized but had to be identified.

The distinction between immediate recognition and identification corresponds closely to the distinction made in the two components theory of recognition (Mandler, 1980) of familiarity and recollection (Yonelinas, 2002). Familiarity, like immediate recognition, is a fast and automatic process. Recollection, akin to identification, lacks this immediacy and is experienced as an effortful search of memory. Despite this patent similarity, there is a difference in focus. Memory research focuses uniquely on recovering the memory. Mental files, on the other hand, draw our attention to a quite different problem: What to make of the test item once it has been conceptualized as a different entity than the one to be remembered? In case of successful retrieval, the test item has to be identified with the remembered item, a process of linking coreferential files. To our knowledge, memory research has overlooked this process.

Counterfactual Thinking

The data we have just reviewed suggest that our mind makes a clear distinction between cases of social cognition where perspective differences are essential and those where they don't matter. Use of a theory of mind would not differ greatly for these two cases. For teleology, a deep-seated change from basic teleology to "teleology-in-perspective" is asked for, which implies that reasoning with false belief requires the ability to reason counterfactually (Perner & Roessler, 2010). Riggs, Peterson, Robinson, and Mitchell (1998) reported a connection for 3- to 5-year-old children, replicated in many studies (review: Rafetseder & Perner, 2018). Rafetseder, Cristi-Vargas, and Perner (2010) and Rafetseder, Schwitalla, and Perner (2013) showed that some counterfactual reasoning tasks are not solved until a few years later. By modifying these tasks so that counterfactual as well as false-belief questions can be asked, Rafetseder and Perner (2018; Rafetseder et al., unpublished) discovered that the false-belief questions became as difficult as the counterfactual questions. These findings suggest that there is a common intellectual component for counterfactual reasoning and reasoning about belief.

Van Hoeck et al. (2014) presented short vignettes, e.g., "Jonas takes a shower, the pizza arrives, Marion takes Jonas's wallet from his trouser pocket to pay, and leaves it on the table." Then, one of three questions was asked, "Where does Jonas expect his wallet to be?" for beliefs (FB), "If Jonas had laid out money for the pizza, where would his wallet be?" for counterfactuals (CF), and "When Jonas gets out of the shower, then where is his wallet?" for control (C). Contrasting each question against the control and testing for the conjunction of contrasts (FB > C and CF > C), three significant areas emerged. One of them is in left IPL right where the identity vs. no-identity contrast for statements, equations, and faces intersect (Fig. 8, yellow sphere). In light of the reviewed data, this coincidence confirms the developmental conclusion that false-belief and counterfactual reasoning share the cognitive mechanism for linking coreferential files, which is implemented in the left IPL.

Conclusion

We made a plea for rethinking the standard approach to social cognition as a theory of mind by bringing together several topics.

1. *Basic Framework*: We started with the question of whether our understanding of the mind consists of a theory (“theory” theory) is based on simulation, or is based on teleology.
2. *Perspective Differences*: Data from development and brain imaging show a strong distinction between cases where perspective differences are essential and those where they can be safely ignored. This distinction shows in the age at which children master such cases and in the involvement of a quite circumscribed cerebral region in the left IPL.
3. *Perspective and Identity*: The reviewed data also show that dealing with perspective differences goes hand in hand with identity judgments, as apparent in synchronized development and a common brain region. Mental files theory helps to explain this curious finding. In general, two files make one think of two objects, even when the files happen to co-refer to the same object. Identity information is to correct this, and the two files need to be *linked* so that one treats them as pertaining to a single object. This model can be extended to capture difference of perspective by having coreferential files for objects, e.g., a regular file with information pertaining to one’s own perspective and a vicarious file pertaining to another person’s perspective. To avoid the impression that oneself and the other are thinking about different objects, the files have to be *linked*. The ability to *link* files is the common denominator. Its emergence at a particular age accounts for the developmental synchronicity, and its special processing requirements account for the activation of a common expert brain region.
4. *Counterfactuals*: Imaging data suggest that the processes in left IPL are shared by thinking about beliefs as well as counterfactual considerations. Moreover, the common region overlaps with the region for linking coreferential files. Developmental data show that the answer pattern to counterfactual questions is reflected in answers to the false-belief question over a broad age range.

Conclusion: The facts under (2) that a variety of perspective and identity problems are not mastered before a certain age (≈4 years) and that they all activate a common brain region (⊂ left IPL) and the fact (4) that counterfactual reasoning relates to reasoning with false belief provide an answer to our question in (1). These data speak for teleology and teleology-in-perspective as our basic model of social cognition.

Collateral Bonus: The left IPL counts as one of the minor “hubs” in the brain (Humphreys & Lambon Ralph, 2017). Hubs interconnect different networks

(Utevsky, Smith, & Huettel, 2014; van den Heuvel & Sporns, 2011) and respond to the different contents processed in these networks. Their cognitive function is still not fully understood (Humphreys & Lambon Ralph, 2014). The reviewed data suggest an answer for, at least, the left IPL. It is a specialist for linking coreferential files, which is needed in many networks (language, numerical, social cognition, person recognition, …). It, thus, receives input from very different domains and is linked strongly to other hubs to exchange information with these different domains.

Acknowledgments The research on which this work is based was funded by the Austrian Science Fund's Doctoral College (DK W 1233-G17) "Imaging the Mind: Connectivity and Higher Cognitive Function," the individual project FWF I 3518-G24 as part of the DACH collaborative project, "The structure and development of understanding actions and reasons,", an Erwin Schroedinger Fellowship FWF-J4009-B27 and a MarieSkłodowska-Curie Individual Fellowship MSCA-IF 844734. Michael Gilead's many helpful improvement suggestions were gratefully implemented.

References

Aichhorn, M., Perner, J., Weiss, B., Kronbichler, M., Staffen, W., & Ladurner, G. (2009). Temporoparietal junction activity in theory-of-mind tasks: Falseness, beliefs, or attention. *Journal of Cognitive Neuroscience, 21*(6), 1179–1192. https://doi.org/10.1162/jocn.2009.21082

Alvarez, M. (2018). Reasons for action, acting for reasons, and rationality. *Synthese, 195*, 3293–3310.

Anderson, J., & Hastie, R. (1974). Individuation and reference in memory: Proper names and definite descriptions. *Cognitive Psychology, 6*, 495–514. https://doi.org/10.1016/0010-0285(74)90023-1

Anscombe, G. E. (1957). *Intention*. Cambridge: Cambridge University Press.

Arora, A., Schurz, M., & Perner, J. (2017). Systematic comparison of brain imaging meta-analyses of ToM with vPT. *BioMed Research International, 2017*, 1–12. https://doi.org/10.1155/2017/6875850

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14*(3), 110–118. https://doi.org/10.1016/j.tics.2009.12.006

Capgras, J., & Reboul-Lachaux, J. (1923). The illusion of 'doubles' in a chronic systematized delusion (Illusion des «sosies» dans un délire systématisé chronique). *Bulletin de La Société Clinique de Médicine Mentale, 2*, 6–16.

Carey, S. (2009). *The origins of concepts*. New York, NY: Oxford University Press.

Charles, D. (2012). Teleological causation. In C. Shields (Ed.), *The Oxford handbook of Aristotle*. Oxford: Oxford University Press.

Clark, E. V. (1997). Conceptual perspective and lexical choice in acquisition. *Cognition, 64*, 1–37. https://doi.org/10.1016/S0010-0277(97)00010-3

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development, 9*, 377–397. https://doi.org/10.1016/0885-2014(94)90012-4

Darby, R. R., & Caplan, D. (2016). 'Cat-gras' delusion: A unique misidentification syndrome and a novel explanation. *Neurocase, 22*(2), 251–256. https://doi.org/10.1080/13554794.2015.1136335

Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy, 60*, 685–700. https://doi.org/10.2307/2023177

Ellis, H. D., & Lewis, M. B. (2001). Capgras delusion: A window on face recognition. *Trends in Cognitive Sciences, 5*, 149–156. https://doi.org/10.1016/S1364-6613(00)01620-X

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1 - Level 2 distinction. *Developmental Psychology, 17*, 99–103. https://doi.org/10.1037/0012-1649.17.1.99

Frege, G. (1892). On sense and reference. In P. Geach & M. Black (Eds.), *Philosophical writings of Gottlob Frege* (Vol. 1, pp. 56–78). Oxford: Basil Blackwell.

Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia, 45*(1), 32–41. https://doi.org/10.1016/j.neuropsychologia.2006.04.015

Gobbini, M. I., Koralek, A. C., Bryan, R. E., Montgomery, K. J., & Haxby, J. V. (2007). Two takes on the social brain: A comparison of theory of mind tasks. *Journal of Cognitive Neuroscience, 19*, 1803–1814. https://doi.org/10.1162/jocn.2007.19.11.1803

Gopnik, A., & Meltzoff, A. N. (1997). *Word, thoughts, and theories. A Bradford Book*. Cambridge, MA: MIT Press.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language, 7*, 145–171. https://doi.org/10.1111/j.1468-0017.1992.tb00202.x

Gordon, R. M. (1986). Folk psychology as simulation. *Mind & Language, 1*, 158–171. https://doi.org/10.1111/j.1468-0017.1986.tb00324.x

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences, 2*(12), 493–501.

Hamilton, A. F., Brindley, R., & Frith, U. (2009). Visual perspective taking impairment in children with autistic spectrum disorder. *Cognition, 113*(1), 37–44. https://doi.org/10.1016/j.cognition.2009.07.007

Heal, J. (1986). Replication and functionalism. In J. Butterfield (Ed.), *Language, mind, and logic* (Vol. 1, pp. 135–150). Cambridge: Cambridge University Press.

Hirstein, W. (2010). The misidentification syndromes as mindreading disorders. *Cognitive Neuropsychiatry, 15*(1–3), 233–260. https://doi.org/10.1080/13546800903414891

Hirstein, W., & Ramachandran, V. S. (1997). Capgras syndrome: A novel probe for understanding the neural representation of the identity and familiarity of persons. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 264*(1380), 437–444. https://doi.org/10.1098/rspb.1997.0062

Humphreys, G. F., & Lambon Ralph, M. A. (2014). Fusion and fission of cognitive functions in the human parietal cortex. *Cerebral Cortex, 25*(10), 3547–3560. https://doi.org/10.1093/cercor/bhu198

Humphreys, G. F., & Lambon Ralph, M. A. (2017). Mapping domain-selective and counterpointed domain-general higher cognitive functions in the lateral parietal cortex: Evidence from fMRI comparisons of difficulty-varying semantic versus visuo-spatial tasks, and functional connectivity analyses. *Cerebral Cortex, 27*(8), 4199–4212. https://doi.org/10.1093/cercor/bhx107

Iao, L. S., & Leekam, S. R. (2014). Nonspecificity and theory of mind: New evidence from a nonverbal false-sign task and children with autism spectrum disorders. *Journal of Experimental Child Psychology, 122C*, 1–20. https://doi.org/10.1016/j.jecp.2013.11.017

Jackson, F., & Braddon-Mitchell, D. (1998). Belief as a map by which we steer. In E. Craig (Ed.), *Routledge encyclopedia of philosophy*. London; New York, NY: Taylor & Francis.

Kulke, L., & Rakoczy, H. (2018). Implicit theory of mind–An overview of current replications and non-replications. *Data in Brief, 16*, 101–104.

Leekam, S., & Perner, J. (1991). Does the autistic child have a metarepresentational deficit? *Cognition, 40*, 203–218. https://doi.org/10.1016/0010-0277(91)90025-Y

Leekam, S., Perner, J., Healey, L., & Sewell, C. (2008). False signs and the non-specificity of theory of mind: Evidence that preschoolers have general difficulties in understanding representations. *British Journal of Developmental Psychology, 26*(4), 485–497. https://doi.org/10.1348/026151007X260154

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87*, 252–271. https://doi.org/10.1037/0033-295X.87.3.252

Masangkay, Z. S., McCluskey, K. A., McIntyre, C. W., Sims-Knight, J., Vaughn, B. E., & Flavell, J. H. (1974). The early development of inferences about the visual percepts of others. *Child Development, 45*, 357–366. https://doi.org/10.2307/1127956

Molenberghs, P., Johnson, H., Henry, J. D., & Mattingley, J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews, 65*, 276–291.

Moll, H., & Tomasello, M. (2006). Level I perspective-taking at 24 months of age. *British Journal of Developmental Psychology, 24*, 603–613. https://doi.org/10.1348/026151005X55370

Neumann, A., Thoermer, C., & Sodian, B. (2008). *False belief understanding in 18-month-olds' anticipatory looking behavior: An eye-tracking study*. Paper presented at the XXIX International Congress of Psychology, Berlin, Germany.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*, 255–258. https://doi.org/10.1126/science.1107621

Parkin, L. J. (1994). *Children's understanding of misrepresentation*. Unpublished doctoral dissertation, University of Sussex.

Perner, J. (2016). Referential and cooperative Bias: In defence of an implicit theory of mind. *Commentary for Symposium on Katharina Helming, Brent Strickland, and Pierre Jacob's "Solving the puzzle about early belief-ascription"*. Retrieved from http://philosophyofbrains.com/

Perner, J., Aichhorn, M., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Thinking of mental and other representations: The roles of left and right temporo-parietal junction. *Social Neuroscience, 1*, 245–259. https://doi.org/10.1080/17470910600989896

Perner, J., & Esken, F. (2015). Evolution of human cooperation in Homo Heidelbergensis: teleology versus mentalism. *Developmental Review, 38*, 69–88. https://doi.org/10.1016/j.dr.2015.07.005

Perner, J., & Leekam, S. (2008). The curious incident of the photo that was accused of being false: Issues of domain specificity in development, autism, and brain imaging. *The Quarterly Journal of Experimental Psychology, 61*, 76–89. https://doi.org/10.1080/17470210701508756

Perner, J., Mauer, M. C., & Hildenbrand, M. (2011). Identity: Key to children's understanding of belief. *Science (New York, N.Y.), 333*(6041), 474–477. https://doi.org/10.1126/science.1201216

Perner, J., & Roessler, J. (2010). Teleology and causal reasoning in children's theory of mind. In J. Aguilar & A. Buckareff (Eds.), *Causing human action: New perspectives on the causal theory of action. A Bradford Book* (pp. 199–228). Cambridge, MA: MIT Press.

Piaget, J., & Inhelder, B. (1948). *The child's conception of space*. London: Routledge; Kegan Paul.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences, 1*, 516–526.

Rafetseder, E., Cristi-Vargas, R., & Perner, J. (2010). Counterfactual reasoning: Developing a sense of 'nearest possible world'. *Child Development, 81*(1), 376–389.

Rafetseder, E., O'Brian, C., Leahy, B., & Perner, J. (2018). *Extended difficulties with counterfactuals persist in reasoning with false beliefs: Evidence for teleology-in-perspective*. Unpublished manuscript, Department of Psychology, University of Stirling, UK.

Rafetseder, E., & Perner, J. (2018). Belief and counterfactuality: A teleological theory of belief attribution. *Zeitschrift für Psychologie, 226*(2), 110–121. https://doi.org/10.1027/2151-2604/a000327

Rafetseder, E., Schwitalla, M., & Perner, J. (2013). Counterfactual reasoning: From childhood to adulthood. *Journal of Experimental Child Psychology, 114*(3), 389–404. https://doi.org/10.1016/j.jecp.2012.10.010

Ramsey, F. P. (1931). *The foundations of mathematics*. London: Kegan Paul.

Raz, J. (1999). Explaining normativity: On rationality and the justification of reason. *Ratio, XII*, 354–379.

Recanati, F. (2012). *Mental files*. Oxford: Oxford University Press.

Riggs, K. J., Peterson, D. M., Robinson, E. J., & Mitchell, P. (1998). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development, 13*(1), 73–90. https://doi.org/10.1016/S0885-2014(98)90021-1

Roessler, J., & Perner, J. (2013). Teleology: Belief as perspective. In S. Baron-Cohen, M. Lombardo, & H. Tager-Flusberg (Eds.), *UOM-3: Understanding other minds* (3rd ed., pp. 35–50). Oxford: Oxford University Press.

Ruffman, T., Garnham, W., Import, A., & Connolly, D. (2001). Does eye gaze indicate implicit knowledge of false belief? Charting transitions in knowledge. *Journal of Experimental Child Psychology, 80*, 201–224. https://doi.org/10.1006/jecp.2001.2633

Sarnecka, B. W., & Wright, C. E. (2013). The idea of an exact number: Children's understanding of cardinality and equinumerosity. *Cognitive Science, 37*(8), 1493–1506. https://doi.org/10.1111/cogs.12043

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporoparietal junction in 'theory of mind'. *NeuroImage, 19*, 1835–1842. https://doi.org/10.1016/S1053-8119(03)00230-1

Scanlon, T. (1998). *What we owe to each other*. Cambridge, MA: Harvard University Press.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences, 19*(2), 65–72. https://doi.org/10.1016/j.tics.2014.11.007

Schurz, M. (2015, June). *Discovering the neural link between theory of mind and visual perspective taking: Issues of spontaneity and domain-specificity*. Presented at the 41st Psychology and Brain Meeting, Frankfurt a.M., Germany.

Schurz, M., Aichhorn, M., Martin, A., & Perner, J. (2013). Common brain areas engaged in false belief reasoning and visual perspective taking: A meta-analysis of functional brain imaging studies. *Frontiers in Human Neuroscience, 7*, 712. https://doi.org/10.3389/fnhum.2013.00712

Schurz, M., Kronbichler, M., Weissengruber, S., Surtees, A., Samson, D., & Perner, J. (2015). Clarifying the role of theory of mind areas during visual perspective taking: Issues of spontaneity and domain-specificity. *NeuroImage, 117*, 386–396.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews, 42*, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009

Schurz, M., Tholen, M. G., Perner, J., Mars, R. B., & Sallet, J. (2017). Specifying the brain anatomy underlying temporo-parietal junction activations for theory of mind: A review using probabilistic atlases from different imaging modalities. *Human Brain Mapping, 38*(9), 4788–4805.

Setoh, P., Scott, R. M., & Baillargeon, R. (2016). Two-and-a-half-year-olds succeed at a traditional false-belief task with reduced processing demands. *Proceedings of the National Academy of Sciences, 113*, 13360–13365.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science, 18*, 580–586. https://doi.org/10.1111/j.1467-9280.2007.01943.x

Surtees, A., Apperly, I., & Samson, D. (2013). Similarities and differences in visual and spatial perspective-taking processes. *Cognition, 129*(2), 426–438. https://doi.org/10.1016/j.cognition.2013.06.008

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological science, 18*(7), 587–592.

Thoermer, C., Sodian, B., Vuori, M., Perst, H., & Kristen, S. (2012). Continuity from an implicit to an explicit understanding of false belief from infancy to preschool age: False belief from infancy to preschool. *British Journal of Developmental Psychology, 30*(1), 172–187. https://doi.org/10.1111/j.2044-835X.2011.02067.x

Tholen, M. G., Schurz, M., & Perner, J. (2019). The role of the IPL in person identification. *Neuropsychologia, 129*, 164–170. https://doi.org/10.1016/j.neuropsychologia.2019.03.019

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., … Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage, 15*(1), 273–289.

Utevsky, A. V., Smith, D. V., & Huettel, S. A. (2014). Precuneus is a functional core of the default-mode network. *The Journal of Neuroscience, 34*(3), 932–940. https://doi.org/10.1523/JNEUROSCI.4227-13.2014 van den Heuvel, M. P., & Sporns, O. (2011). Rich-club organization of the human connectome. *Journal of Neuroscience, 31*(44), 15775–15786. https://doi.org/10.1523/JNEUROSCI.3539-11.2011

Van Hoeck, N., Begtas, E., Steen, J., Kestemont, J., Vandekerckhove, M., & Van Overwalle, F. (2014). False belief and counterfactual reasoning in a social environment. *NeuroImage, 90*, 315–325. https://doi.org/10.1016/j.neuroimage.2013.12.043

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30*(3), 829–858. https://doi.org/10.1002/hbm.20547

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory of mind development: The truth about false belief. *Child Development, 72*, 655–684. https://doi.org/10.1111/1467-8624.00304

Wilkinson, S. (2016). A mental files approach to delusional misidentification. *Review of Philosophy and Psychology, 7*(2), 389–404. https://doi.org/10.1007/s13164-015-0260-5

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441–517.

Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and 'false' photographs. *Cognition, 35*, 41–68. https://doi.org/10.1016/0010-0277(90)90036-J

The Organization of Social Knowledge Is Tuned for Prediction

Mark A. Thornton and Diana I. Tamir

The social world requires people to make highly consequential predictions. People need to predict whether new acquaintances will become a friend or foe (Cuddy, Fiske, & Glick, 2008), how old friends will respond to constructive criticism, or how long the boss will be angry before approaching her for a favor. Whether in cooperation or competition, successful social interaction requires people to anticipate others' future thoughts, feelings, and actions, and prepare their own actions accordingly. Social predictions are among the most common predictions a person must make because people spend so much time with other people (United States Bureau of Labor Statistics, 2003). Yet despite the importance of social prediction, researchers have only just scratched the surface of the predictive social mind. Here we consider recent research that is starting to reveal how people glimpse the social future.

Research on nonsocial prediction suggests that the brain is built to make predictions. It does not passively perceive the world around it and then react accordingly. Instead, people make reflexive predictions across multiple domains (Hohwy, Roepstorff, & Friston, 2008; Rao & Ballard, 1999; Vuust, Ostergaard, Pallesen, Bailey, & Roepstorff, 2009). When processing language, for example, people use the beginning of a sentence to predict the end of that sandwich. When watching a ball thrown into the air, people reflexively make a prediction about its eventual downward trajectory. However, the social world poses unique challenges to people's well-honed predictive capacities. Humans are not billiard balls: people are probabilistic beings, moved to action by the unseen forces of thoughts and feelings.

M. A. Thornton (✉)
Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA

D. I. Tamir
Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA
e-mail: mark.a.thornton@dartmouth.edu; dtamir@princeton.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_14

How do people represent these invisible mental states, and use them to make predictions?

A recent theoretical framework for social cognition (Tamir & Thornton, 2018) proposes a simple answer to this question (Fig. 1). This multilayered framework of

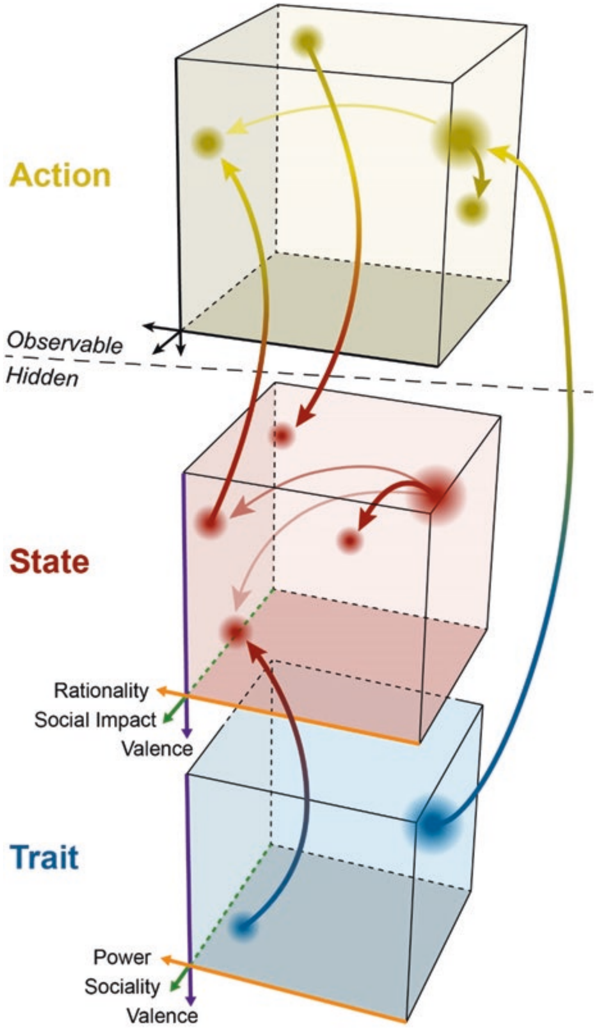


Fig. 1 The multilayer model of predictive social cognition. Three layers of social knowledge—observable actions and hidden mental states and traits—are each organized into low-dimensional maps by psychological dimensions such as valence and power. Transitions between or within layers (arrows) decrease in probability with distance. Short hops between adjacent points (e.g., happiness and gratitude) are more likely than long treks between distant points (e.g., sleeping and hiking). This organization of social knowledge provides both for parsimonious representation of the complexities of the social world, and for accurate, automatic prediction of the social future. (Reproduced with permission from *Trends in Cognitive Science*)

social cognition helps to explain how people predict others' future states and behaviors in two steps: First, it suggests that the mind organizes social knowledge using conceptual "maps" of social stimuli. These maps allow people to easily track other people's current thoughts, feeling, and actions. Second, it suggests that people track distances and trajectories through these maps to make efficient, automatic social predictions. This framework advances prediction as the central goal of representing social knowledge.

In this chapter, we first describe how people simplify the complexity of the social world using these low-dimensional maps. Next, we discuss how people leverage these maps to make accurate, automatic social predictions. Finally, we offer suggestions for how future research can use this framework productively to model real-world social predictions, and constructively to enhance its explanatory power and scope.

The Organization of Social Knowledge

Humans enjoy extraordinarily rich social lives. Most people know hundreds or even thousands of unique individuals. In a given moment, each one of these individuals may be thinking widely differing thoughts, experiencing different emotions, or performing different actions. This richness makes life interesting and exciting, but it also poses a significant challenge: people must understand others' traits, mental states, and action in order to successfully interact with them. How can the social mind come to grips with the complexity of social life?

The right organizational scheme can bring order to chaos. What kind of scheme does the human mind use to organize social information? We propose that people use a social map. Or rather, that people employ multiple maps, one for each type of social information. Maps organize information by localizing it to particular coordinates on a small set of continuous dimensions. In a geographic map of the United States, the physical location of any city can be described in terms of its dimensions of longitude and latitude. Simply knowing the north-south and east-west coordinates of a city allows you to extract important information about its location from single pair of numbers. For instance, you might learn that a city is located in the Pacific Northwest, and thus that it is nearby to Seattle, WA, but far away from Miami, FL.

Conceptual maps of the social world act in a similar way, by reducing the complexity of social stimuli down to a few essential values. Social neuroscience research has begun to chart the maps that the brain uses to make sense of the social world. This work has revealed the cardinal dimensions that describe three key types of social information: actions, mental states, and traits. These three layers of social information form the core of what one might want to know about a person: what are they doing, how are they feeling, and what kind of person are they? Each layer captures the world at a different timescale. Actions occur at the shortest timescales, from less than a second up to a few hours, depending on their complexity. States might unfold over just a few minutes or persist for several days, depending on if they are more like emotions or moods. Traits are more lasting, or even permanent, ways to describe individual differences among people. To understand and predict other people at any time scale requires us to map out the content of each of these layers of social knowledge.

Mapping the Action Layer

Humans have an incredibly diverse behavioral repertoire. People are capable of engaging in thousands of different actions and activities, ranging from simple motor actions such as reaching and grasping, to complex extended activities such as conducting research or governing a nation. A successful social agent must have a keen understanding of these actions. Recent research (Thornton & Tamir, 2019b) has identified six psychological dimensions (Table 1) that scaffold people's action concepts: the Abstraction, Creation, Tradition, Food, Animacy, and Spiritualism Taxonomy or ACT-FAST. These dimensions originated from a data-driven principal component analysis of verb use in large text corpora and were validated against behavioral judgments.

Like the other social maps we will explore, each dimension of the ACT-FAST carries intrinsic meaning. For example, if one knows that an action is high on the animacy dimension, then one knows that it is an action that tends to be performed by living agents such as humans and animals, rather than an act of nature or a machine. In this way, this conceptual map of the social world both offers the functionality of a physical map—representing the distances between difference locations—and also implies rich knowledge about each of those locations.

Together, the six ACT-FAST dimensions explain much of how people think about actions. Knowing an action's coordinates on these dimensions can robustly predict: (1) *who* does an action, in terms of traits, (2) *why* one does an action, in terms of approach and avoidance motivations, (3) *when* one does an action, in terms of time of day, (4) *where* one does an action, in terms of outdoor versus indoor, and if indoor, public versus private, and (5) *how* one performs an action, in terms of body parts involved and mental or physical effort required. ACT-FAST can also explain patterns of natural language use, such as which verbs tend to co-occur with

Table 1 Dimensions of the FAACTS action taxonomy

Dimension	Pole 1	Pole 2	Examples
Food	Food	Nonfood	Bake, fry vs. detain, testify
Abstraction	Abstract/social	Concrete/physical	Govern, refute vs. drip, peel
Animacy	Animate	Mechanical	Meow, floss vs. contain, extract
Creation	Creation	Crime	Film, sing vs. prosecute, testify
Tradition	Tradition	Innovation	Cook, decorate vs. emit, encrypt
Spiritualism	Work	Worship	Fax, haggle vs. foretell, ascend

which nouns. In addition to answering these important psychological and linguistic questions, ACT-FAST can explain brain activity across a wide set of cortical regions implicated in action representation. Together, these findings suggest that ACT-FAST provide a useful, and biologically plausible map of how people organize knowledge about other people's actions.

Mapping the State Layer

In addition to paying attention to others' observable actions, successful social agents also attend to the hidden drivers of actions: mental states. These states must be inferred from indirect cues, such as facial expression and tone of voice, but once known, others' thoughts and feelings can serve as powerful predictors of their behavior. People often share the same intuitions about the predictive power of mental states: angry people aggress, tired people rest, and happy people celebrate. Failing to attend to others' mental states could lead to embarrassing faux pas at best, or to serious danger at worst. To avoid such pitfalls, we need a map of the state layer.

A number of established theories propose dimensions that organize mental states. For example, the circumplex model of affect offers a map with two dimensions, valence and arousal (Russell, 1980); or the distinction between emotional and rational states, prominent in many modern dual-process theories (Evans & Stanovich, 2013; Kahneman, 2003). Recent neuroimaging work (Tamir, Thornton, Contreras, & Mitchell, 2016) used PCA to synthesize the 16 dimensions that comprised these existing theories, and then validated the resulting components in terms of their ability to predict neural representations of mental states. This work identified three dimensions that structure the brain's map of others' mental states (Table 2). The first dimension, valence, captures whether others are feeling good or bad. Knowing the valence of a person's mental state could help people avoid harm from those in negative states like rage, and enjoy pleasant, constructive social interactions with those in more positive moods. The second dimension on this map, social impact, captures which mental states would dispose others to engage in social interactions. Highly impactful states, whether good or bad, are more likely to affect one's life. The final dimension, rationality, captures whether others are likely to act in a calm, deliberate, well-thought-out way, or react instinctively or rashly.

Valence, social impact, and rationality together comprise the 3d Mind Model. This model can explain over 80% of the variance in neural representations of mental states. That is, it provides a near-complete map of the mental state layer (Thornton

Table 2 Dimensions of the mental state representation

Dimension	Pole 1	Pole 2	Examples
Valence	Positive	Negative	Ecstasy, peacefulness vs. rage, sadness
Social impact	Impactful	Unimpactful	Envy, love vs. exhaustion, self-pity
Rationality	Rational	Emotional	Thinking, planning vs. joy, lust

& Tamir, 2020a). Moreover, this map remains robust across different ways of perceiving mental states: similar dimensions emerge across modalities, regardless of whether people reflect on mental state-related scenarios presented as images or as text (Weaverdyck, Thornton, & Tamir, 2020). People also apply similar maps to thinking about their own minds, as opposed to the minds of others. While the structure of the map doesn't change across targets, the resolution does: when people think about their own mental states, they pore over a highly detailed, richly annotated map (Thornton, Weaverdyck, Mildner, & Tamir, 2019). In contrast, when people think about the states of others, the resolution is much less fine-grained. This difference likely results from both the quality and quantity of information one has about their own minds, in contrast to the minds of others.

Mapping the Trait Layer

Although people may apply similar maps to the mental states of different people, individuals do differ in socially important ways. Enduring individual differences between people are known as traits. Compared with actions and mental states, traits are relatively permanent fixtures of an individual, changing slowly across a lifetime, if at all (Roberts & Mroczek, 2008; Srivastava, John, Gosling, & Potter, 2003). Knowing where a person places on trait dimensions can help people to make predictions about their likely states or actions. For example, if you know that someone is highly trustworthy, you can predict that they will not steal from you; if you know someone is highly social, you can predict that they might feel excited at a party. Traits thus help people to make nuanced predictions about people, across situations.

There are multiple existing dimensional maps of traits, including the Five Factor model (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) of personality (Goldberg, 1990; McCrae & Costa, 1987), and the stereotype content model (Fiske, Cuddy, Glick, & Xu, 2002) consisting of warmth and competence. Recent neuroimaging research synthesized several of the most prominent trait theories in the literature (Thornton & Mitchell, 2018). In this work, a three-dimensional model—consisting of power, valence, and sociality—provided the best explanation for patterns of brain activity elicited by thinking about a large set of public figures. Knowing where a person resides on these dimensions can inform social judgments about them. Power indicates whether another person is dominant and competent, and thus, capable of enacting their will. Valence indicates whether that person is warm and trustworthy, and thus, likely to help or harm. Finally, sociality reflects whether a person is extraverted, and thus, likely to engage in the first place. Together, these dimensions provide a near-complete map of the trait space: Power, valence, and sociality explain more than two-thirds of reliable neural activity associated with making inferences about other people. This model outperforms even the Big 5 personality traits in explaining neural representations of other people. However, it is important to note that the Big 5 are ostensibly a model of the reality of traits, whereas the three-dimensional model instead aims to explain the perception of traits. Moreover, it is an open question whether people continue to apply any map of trait space for personally familiar others (Thornton & Mitchell, 2017). When people interact regularly, they may instead draw on the character of their relationship itself.

Overlapping Social Maps

The maps our mind makes of the social world rely on a similar set of brain regions. Both mental state and trait representation engage regions such as the medial prefrontal and parietal cortices, and the superior temporal sulcus extending from the temporoparietal junction forward to the anterior temporal lobe (Tamir et al., 2016; Thornton & Mitchell, 2018). These regions form the social brain network, a set of brain regions reliably activated by a wide range of social stimuli (Mitchell, 2008; Van Overwalle & Baetens, 2009). However, mental states and traits share more than just a gross anatomical similarity—they share a common neural code. Both maps include a dimension reflecting valence; both include a dimension reflecting sociality; and both include a dimension reflecting competence, dominance, and agency. This allows each dimension to be decoded from brain activity across domains (Thornton & Mitchell, 2018). This shared code hints at a deep connection between the way people think about others' momentary and enduring mental properties—potentially bridging the traditional divide between traits and states.

Indeed, in recent research, we found that when our brain represents a person, it seems to do so by keeping track of the mental states that person habitually experiences (Thornton, Weaverdyck, & Tamir, 2019a). For example, if a politician is habitually in bad moods—grouchy, short-tempered, stubborn—one may form the impression that the politician has a negative disposition, at the trait level. Correspondingly, our results indicate that the pattern elicited by thinking about such a person could be reconstructed by adding together generic representations of grouchiness, short-temper, and stubbornness. This finding suggests a simple mechanism for impression formation—counting perceived mental states. Moreover, this process could later be reversed to make predictions about states based on people's dispositions. Thus, although traits and states are typically thought of as separate, the trait and state dimensions described above reflect parallel concepts.

However, not all socially relevant information occupies the same neural territory. In recent research, we and others have found that actions are represented in quite different portions of the brain than those involved in theory of mind (Tarhan & Konkle, 2019; Thornton & Tamir, 2020b). High-level visual regions within the dorsal and ventral paths—areas rarely implicated in social cognition per se—appear to play an important role in representing what others are doing. Understanding more about how maps of the social world relate—both in conceptual space and the physical territory of the brain—is a high priority for future research.

The Predictive Social Brain

Conceptual maps of actions, mental states, and traits allow people to organize their knowledge of the social world. However, the social world is not static; it is constantly in flux. At one moment, a person may be having fun dancing with their friends, but after a few hours of this energetic state, they may find themselves feeling exhausted. This mental state of exhaustion may in turn lead to a new activity, like resting, and so forth. Navigating the social world requires people to anticipate such transitions. Fortunately, the dimensional maps of social knowledge described above offer insight into these dynamics: knowing "where" a person is on a social map can tell you a great deal about where that person will "go" in the future.

Social Dimensions Scaffold Social Predictions

How can the static maps described above be used to predict the social future? The key assumption that licenses such predictions is that the proximity between points on the map reflects the likelihood of moving between those points. This assumption holds for geographic maps: on average, people tend to travel locally—e.g., to the store, work, school, or gym—with relatively high frequency. People travel to more distant locales—another city, country, or continent—much more rarely. Consequently, if you know another person's current location on a map, you can accurately predict that their next destination is going to be close by.

We propose that people apply an analogous algorithm to make social predictions using conceptual maps. The coordinates on these maps represent actions, mental states, and personality traits rather than physical locations, but the logic of distance-based prediction still applies. Simply put, people are more likely to transition from one location within a layer to a nearby location than they are to transition to faraway locations. For example, if a person is currently feeling pride, a positive emotion, one might guess that she is much more likely to feel happy next than to feel sad. In this way, one could predict the likely trajectory of a person through a layer simply by knowing that person's current coordinates. If so, then the multilayer structure mapped above would provide the foundation for understanding the dynamics of the social world.

Behavioral research suggests that people do predict the likelihood of transitions between states based on the proximity between those states on the dimensions of the 3d Mind Model—valence, social impact, and rationality. To demonstrate this, we first elicited people's intuitions for how states transition from one to the next (Thornton & Tamir, 2017). For example, a participant might be told that a person is currently feeling "excitement" and be asked to rate the likelihood that that person will next experience "sleepiness" from 0% to 100%. Each state was mapped according to its location on the dimensions of rationality, social impact, and valence so that we could calculate the "distance" between each pair of states. Across two studies,

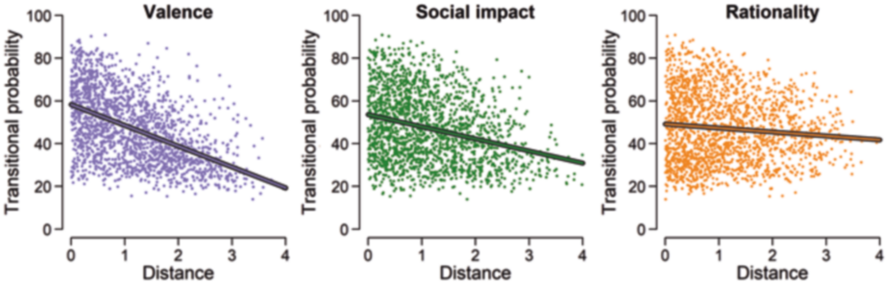


Fig. 2 Dimensional proximity predicts transitional probabilities between states. Each point on the scatter plots represents a transition from one mental state to another. The *x*-axis indicates the absolute distance between those states on each mental state dimension. The *y*-axis indicates the predicted transitional probability from one state to the other. The further two states are on each dimension, the less people expect a transition from one to the next. People are also less likely to actually transition between distant states. These state dimensions explain much of the accuracy of predicted transitional probabilities (Thornton & Tamir, 2017)

proximity on the dimensions was positively associated with participants' transitional probability ratings. The closer states were to each other on any of these mental state dimensions, the more likely participants judged the transitions between them (Fig. 2).

In the same way, people use conceptual proximity on the ACT-FAST dimensions to predict others' actions (Thornton & Tamir, 2019a). Across five preregistered behavioral experiments, participants rated the likelihood that a person currently engaged in one action to next engage in another. For instance, how likely would it be for someone currently "dancing" to next "rest"? Proximities between actions on the ACT-FAST action dimensions reliably predict people's transitional probability ratings. As with states, this result suggests that people may draw upon their map of the action layer to make predictions about others' likely future actions, based on their current actions.

Social Dimensions Describe Real Social Dynamics

In the previous section, we described evidence that people use proximity within a social map to make predictions about others' mental states and actions. However, it would only make sense to use social maps to make predictions if the social dimensions describe actual social dynamics. That is, one should use proximity on the dimension of valence to predict mental state transitions if and only if valence in fact describes regularities in the mental state transitions that others actually experience. As part of the investigations described in the previous section, we used experience sampling data and other real-world data to measure actual state and action transitions. This allowed for an estimation of the actual transitional probabilities between pairs of mental states (Thornton & Tamir, 2017) and pairs of actions (Thornton & Tamir, 2019a).

Distance on the mental state map predicted actual mental state transitions. The further away two states were on any dimension, the less likely people were to transition between them. Thus, these dimensions likely serve as a scaffolding for social prediction because they describe experienced social dynamics. Not only were perceived and actual emotion transitions correlated, but these associations could be partially explained by how close the emotions were on the dimensions of mental state representation described above. This indicates that people use their maps of mental state space to accurately predict others' emotion dynamics.

Subsequent research found similar results in the action layer (Thornton & Tamir, 2019a). Across five studies, distance on the action map predicted actual action transitions. The further two actions were on the ACT-FAST dimensions, the less likely people were to transition between them. Moreover, as in the case of mental state representation, people were highly accurate about actual action dynamics, and distance within the action map statistically mediated much of the association between perceived and actual action transitions. This finding adds further weight to the contention that people use their maps of the social world to make accurate social predictions.

The ground truth in all the studies above reflects transitional probabilities aggregated across many people or datasets. These data demonstrate that people make accurate predictions about a "generic" other. More recent data suggest that people can likewise make accurate predictions about both specific people and relevant social groups (Zhao, Thornton, & Tamir, 2018). For example, in one set of studies, we asked undergraduates to make predictions about a specific other—either a close friend, or their current roommate—as well as their undergraduate community in general. In all cases, people were able to accurately and specifically predict their friend, their roommate, and their community. This indicates that people may have highly accurate models of state dynamics in general, and that they can also tailor these models to make predictions about the individuals in their lives.

People Make Social Predictions Automatically

People can predict the social future with such high fidelity because prediction is built into the way that people represent social knowledge. In recent research from the action domain, we found that while people watch a movie, their brains spontaneously encode the actions they perceive on the ACT-FAST dimensions, and these ACT-FAST coordinates predict actions later in the movie (Thornton & Tamir, 2020b). That is, a participant's brain activity at a given moment in time automatically predicts the actions which are actually likely to occur later on. For example, if participants saw a person "running," then they would encode this action on ACT-FAST—attributing to this act a high degree of animacy. This would, in turn, accurately predict that they were likely to see other high animacy actions in the near future. Merely by encoding perceived actions on an appropriate set of dimensions—dimensions which through real actions flow smoothly—the brain thus automatically predicts likely future actions.

Similar evidence for the automaticity of social prediction comes from the domain of mental states. Whenever someone thinks about a mental state, they do not think about that mental state in isolation; they also spontaneously think about likely future states. For example, when one observes a friend experiencing pride, they can accurately predict that the friend will soon feel happy because the representation of pride incorporates the representation of happiness. Neuroimaging work has provided a unique source of evidence that this is the case: neural patterns associated with a mental state currently under consideration literally resembled patterns of likely future states (Thornton, Weaverdyck, & Tamir, 2019b). The more likely one state is to transition to another state, the more similar the neural patterns that represent them. Importantly, this work also showed that transition predictions, and not simply similarity, drove this neural finding. Even though similarity and transition likelihood are highly intertwined concepts, multiple lines of evidence suggest that transitions, and not similarity, may be primary in defining the conceptual space of mental states.

Indirect neural evidence also supports the automaticity of social prediction. In the same study, repetition suppression (also known as fMRI adaptation) tested how the brain reacted to expected and unexpected sequences of mental states. The principle behind this analysis is that, if the brain is constantly making automatic predictions, perceiving information that violates these predictions should elicit more activity than perceiving prediction-consistent information, since the latter requires recalibration of subsequent predictions. In line with this hypothesis, the study found that seeing states in unexpected sequences elicited more activity in the precuneus than observing predictable state sequences. This finding suggests that this brain region might automatically track errors in mental state predictions and update subsequent predictions accordingly.

People's maps of the social world play a key role in making the brain's social predictions automatic. Since the proximity on dimensions such as rationality, social impact, and valence is associated with transitional probabilities, as described in the previous sections, then merely encoding a state using these dimensions implicitly makes a prediction. That is, simply specifying the location of a state on such dimensions provides an indication of which other states are more or less likely. Supporting this idea, this study also found that proximity on these three dimensions of mental state representation statistically mediated much of the relationship between transitional probabilities and neural pattern similarity (Thornton, Weaverdyck, & Tamir, 2019b). This suggests that part of the reason that neural representations of current states resemble neural representations of likely future states may be because all states are encoded as coordinates within the mental state map.

Conclusion

The brain makes sense of other people's minds by charting conceptual maps of social stimuli, such as actions, mental states, and traits. These maps make the deluge of information from social world more tractable by reducing the complexity of these stimuli down to coordinates on a few essential psychological dimensions, such as valence or animacy. Moreover, these maps allow people to make accurate, automatic predictions about others' actions and mental states. Just knowing where someone is on these maps can tell you a great deal about where they are going next because shorter journeys are more likely than lengthy ones. The short hop from joy to gratitude is far more likely than the arduous trek from delight to despair. The research outlined in this chapter offers insight into how people make sense of other people in order to navigate the choppy waters of everyday social life.

So far, research has focused on how the most basic information conveyed by social maps—the locations of traits, states, and actions—can inform social prediction. However, as with physical maps, social maps can also convey other forms of information. For example, if you look at oceans on many physical maps, you will see small arrows that indicate the direction of prevailing winds and currents. If you dropped a sealed bottle in the ocean at given location, you could use its location, along with knowledge about local water currents to make a precise—and directional—prediction about its future location. The space of mental states likewise has prevailing winds and currents: there are statistical regularities in the trajectories that states and actions follow on their respective maps. For example, people in high energy states are likely to gradually flow toward lower energy states as they tire themselves out. Future research must attempt to map these vector fields in the social domain to further refine the predictive framework we describe here.

The framework for predictive social cognition described in this chapter faces at least four additional challenges (Saxe, 2018). First, the types of social knowledge mapped so far all relate to the person—what they are doing, how they are feeling, and their character. However, one of the most potent drivers of real-world behavior does not dwell within any one person. Instead, the situation, or context, crucially shapes how one will think, feel, and act. The current framework must expand to incorporate the power of the situation and its role in social predictions. Fortunately, in recent years, behavioral research and text analysis have suggested potential maps of the situation layer (Parrigon, Woo, Tay, & Wang, 2017; Rauthmann et al., 2014). Future research may productively test how well such maps explain neural representations of situations, and whether these maps have the same predictive properties as the maps of action and state layers described above.

Second, cultural differences may shape the way people construct maps of social knowledge. Societies differ greatly in the way they perceive emotions, the value they place on different traits, and in the actions that people typically perform (Ching et al., 2014; Gendron, Roberson, van der Vyver, & Barrett, 2014; Tsai, Knutson, & Fung, 2006; Watson-Jones & Legare, 2016). Measuring the generalizability and variability across cultures of the dimensions identified above must be another priority for this research program.

Third, although the model of predictive social cognition described in this chapter has demonstrated its ability to predict real social experience, it faces another major challenge in making these predictions precise. Specifically, this probabilistic framework must incorporate propositional information. For instance, the current framework can describe the properties of "desire" as an abstract mental state, but the meaning of desire can change depending on what one desires. Desiring a cheeseburger and desiring a job are both recognizable forms of desire, but each would predict dramatically different behaviors. Other models of theory of mind—such as Bayesian inverse planning models (Baker, Saxe, & Tenenbaum, 2009)—can deal well with these sorts of propositional problems. However, these types of models do not scale to real-world experience as easily as the current model. Finding ways to unite these models may prove challenging, but a variety of emerging methods, such as word vector embeddings to quantify semantics (Pennington, Socher, & Manning, 2014), may help address these challenges, as similar neural networks have shown to also implicitly represent propositional relations (McCoy, Linzen, Dunbar, & Smolensky, 2018).

The fourth major challenge faced by the predictive model of social cognition is the question of how the mind learns to map the social world. The dimensions of mental state representation arise over the course of development—infants do not start off understanding mental states on all of the dimensions which adults do (Nook, Sasse, Lambert, McLaughlin, & Somerville, 2018). Do children learn new dimensions by observing statistical regularities in emotion dynamics? Perhaps—it is well known that children possess the ability to learn the transitional probabilities between components of speech (Saffran, Aslin, & Newport, 1996), so the same might well be true with respect to other social stimuli. However, it is also possible that children may have "built-in" core knowledge, or inductive biases that help them learn social maps more adeptly than they otherwise might. Future developmental and comparative research, as well as study of machine intelligence, may help to answer this question.

As the model of predictive social cognition described in this chapter becomes more comprehensive and more refined, it holds considerable promise for addressing problems of societal importance. For instance, it may provide a concrete way to quantify abnormal social cognition, such as in Autism Spectrum Disorder, or to track how children learn social concepts over development. Indeed, a theory of "Mind-space"—similar to the mental state and trait layers of our predictive model—has been proposed as a way to gain traction on individual differences in social cognition (Conway, Catmur, & Bird, 2019). Predictive social cognition may also provide a roadmap for enhancing artificial intelligence in the social domain, allowing smart devices to better anticipate people's needs and to interact with people in more natural, human-like ways. Finally, the shortcuts the brain takes to social understanding could reveal the precise sources of harmful social biases and suggest potential approaches to mitigating them. Even optimistically, such development remains many years away, but nonetheless, understanding predictive social cognition holds much promise for the future.

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*(3), 329–349.

Ching, C. M., Church, A. T., Katigbak, M. S., Reyes, J. A. S., Tanaka-Matsumi, J., Takaoka, S., … Rincon, B. C. (2014). The manifestation of traits in everyday behavior and affect: A five-culture study. *Journal of Research in Personality, 48*, 1–16.

Conway, J. R., Catmur, C., & Bird, G. (2019). Understanding individual differences in theory of mind via representation of minds, not mental states. *Psychonomic Bulletin & Review*, 1–15.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology, 40*, 61–149.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241.

Fiske, S., Cuddy, A., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878–902.

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion, 14*(2), 251–262.

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1216–1229.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition, 108*(3), 687–701.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review, 93*, 1449–1475.

McCoy, R. T., Linzen, T., Dunbar, E., & Smolensky, P. (2018). RNNs implicitly implement tensor product representations. *ArXiv, 1812*, 08718.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*(1), 81–90.

Mitchell, J. P. (2008). Contributions of functional neuroimaging to the study of social cognition. *Current Directions in Psychological Science, 17*(2), 142–146.

Nook, E. C., Sasse, S. F., Lambert, H. K., McLaughlin, K. A., & Somerville, L. H. (2018). The nonlinear development of emotion differentiation: Granular emotional experience is low in adolescence. *Psychological Science, 29*(8), 1346–1357.

Parrigon, S., Woo, S. E., Tay, L., & Wang, T. (2017). CAPTION-ing the situation: A lexically-derived taxonomy of psychological situation characteristics. *Journal of Personality and Social Psychology, 112*(4), 642.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87.

Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., … Funder, D. C. (2014). The situational eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology, 107*(4), 677–718.

Roberts, B. W., & Mroczek, D. (2008). Personality trait change in adulthood. *Current Directions in Psychological Science, 17*(1), 31–35.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928.

Saxe, R. (2018). Seeing other minds in 3D. *Trends in Cognitive Sciences, 22*(3), 193–195.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology, 84*(5), 1041–1053.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences, 22*(3), 201–212.

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences, 113*(1), 194–199.

Tarhan, L., & Konkle, T. (2019). Sociality and interaction envelope organize visual action representations. *BioRxiv*, 618272.

Thornton, M. A., & Mitchell, J. P. (2017). Consistent neural activity patterns represent personally familiar people. *Journal of Cognitive Neuroscience, 29*(9), 1583–1594.

Thornton, M. A., & Mitchell, J. P. (2018). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex, 28*(10), 3505–3520.

Thornton, M. A., & Tamir, D. I. (2017). Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences, 114*, 5982.

Thornton, M. A., & Tamir, D. I. (2019a). People accurately predict the transitional probabilities between actions. *PsyArXiv.*

Thornton, M. A., & Tamir, D. I. (2019b). Six dimensions describe action understanding: The ACT-FASTaxonomy. *PsyArXiv*.

Thornton, M. A., & Tamir, D. I. (2020a). People represent mental states in terms of rationality, social impact, and valence: Validating the 3d mind model. *Cortex, 125,* 44–59.

Thornton, M. A., & Tamir, D. I. (2020b). Perceiving actions before they happen: Psychological dimensions scaffold neural action prediction. *Social Cognitive and Affective Neuroscience.*

Thornton, M. A., Weaverdyck, M. E., Mildner, J. N., & Tamir, D. I. (2019). People represent their own mental states more distinctly than those of others. *Nature Communications, 10*(2117).

Thornton, M. A., Weaverdyck, M. E., & Tamir, D. I. (2019a). The brain represents people as the mental states they habitually experience. *Nature Communications, 10*(2291).

Thornton, M. A., Weaverdyck, M. E., & Tamir, D. I. (2019b). The social brain automatically predicts others' future mental states. *Journal of Neuroscience, 39*(1), 140–148.

Tsai, J. L., Knutson, B., & Fung, H. H. (2006). Cultural variation in affect valuation. *Journal of Personality and Social Psychology, 90*(2), 288–307.

United States Bureau of Labor Statistics. (2003). *American time use survey*. Washington, DC: Author.

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage, 48*(3), 564–584.

Vuust, P., Ostergaard, L., Pallesen, K. J., Bailey, C., & Roepstorff, A. (2009). Predictive coding of music–brain responses to rhythmic incongruity. *Cortex, 45*(1), 80–92.

Watson-Jones, R. E., & Legare, C. H. (2016). The social functions of group rituals. *Current Directions in Psychological Science, 25*(1), 42–46.

Weaverdyck, M. E., Thornton, M. A., & Tamir, D. I. (2020). The neural geometry of mental state representation is stable across modalities and targets. Manuscript in preparation

Zhao, Z., Thornton, M. A., & Tamir, D. (2018). Accurate prediction of emotion transitions is associated with social benefits. *PsyArXiv*.

Computational Models of Mentalizing

Bryan González and Luke J. Chang

Introduction

Humans have an incredible ability to effortlessly infer another's mental state. What type of computations facilitates this ability? One of the pioneers of social psychology, Kurt Lewin, attempted to develop a formal psychological framework using mathematical tools such as topology and vector calculus (Lewin, 1936, 1938). He posited that behavioral actions, B, can be described as a function of an individual person, P, acting within an environment, E, $B = f(P, E)$ (Lewin, 1936). An individual's goals and mental states will influence their behavior, but the possible trajectories will ultimately be constrained by the environment. This framework may be useful in understanding how we can predict another person's behavior. Mentalizing describes the psychological process of inferring another person's beliefs, intentions, desires, and feelings. These inferences will likely be influenced by the environment (e.g., what would a reasonable person do in this situation?), as well as the specific actions of the person (e.g., they are heading straight to the food, they must be *hungry*). Since the early days of Lewin, there have been many advances in formalizing the computations of mentalizing. In this chapter, we will provide a brief overview of some of the key concepts and recent developments.

How are we able to effortlessly infer others' beliefs, intentions, desires, and feelings? Information cues about these hidden states can be perceived from many sensory modalities. For example, subtle facial expressions can communicate information about an individual's internal state (Darwin, 1886). However, these signals are noisy and do not necessarily directly correspond to specific internal states. As an example, consider how we might infer if another person is enjoying their meal. After taking a bite of specific dish, the person smiles. If smiling was completely involuntary and only occurred when enjoying an experience, then this

B. González · L. J. Chang (✉)
Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA
e-mail: Bryan.S.Gonzalez.GR@dartmouth.edu; luke.j.chang@dartmouth.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_15

should lead us to infer that the person is enjoying the food. However, perhaps the person knows that we personally prepared the dish and is smiling to politely indicate acknowledgment of our efforts. Multiple internal states can produce the same facial expression (Martin, Rychlowska, Wood, & Niedenthal, 2017). Early neurophysiologists demonstrated that there are distinct neural pathways that independently innervate voluntary and involuntary control of facial expressions such as smiles (Duchenne de Boulogne, 1876; Müri, 2016). Yet, we are able to reliably distinguish voluntary produced, from spontaneously induced smiles, and can also interpret the internal states of others from their prosody of speech (Achim, Guitton, Jackson, Boutin, & Monetta, 2013; Baltaxe, 1991), body posture (Parkinson, Walker, Memmi, & Wheatley, 2017), or their complex volitional actions. This sensitivity to the social environment is so acute that we sometimes even interpret nonhuman objects in the external world as possessing volitional agency with their own internal feelings, beliefs, and desires (Heider & Simmel, 1944). Thus, inferring the meaning behind these social behaviors requires reasoning about the potential latent mental states that cause the observed behavior.

Empirical experiments have been instrumental in improving our understanding of the process of mentalizing beyond philosophical thought experiments. Philosophers have proposed a stringent test for the presence of theory-of-mind (ToM)—the prediction of another person's behavior on the basis of a person's false belief (Bennett, 1978; Dennett, 1978; Harman, 1978; Pylyshyn, 1978). In this view, a *true* belief would be insufficient because it would be impossible to discern whether a person behaves in accordance with reality or in accordance with their own *beliefs* about reality. The "Sally-Ann false-belief test" was designed to meet this challenge (Wimmer & Perner, 1983) and has been well studied across a wide range of ages (Wellman, Cross, & Watson, 2001), species (Call & Tomasello, 1999; Krupenye, Kano, Hirata, Call, & Tomasello, 2016), and special populations (Baron-Cohen, Leslie, & Frith, 1985; Rosenthal, Hutcherson, Adolphs, & Stanley, 2019). More recently, the neural basis of mentalizing has been studied when reading stories involving mental states (Saxe & Kanwisher, 2003/2008), making inferences about others' preferences (Jenkins, Macrae, & Mitchell, 2008; Mitchell, Heatherton, & Macrae, 2002), watching movies that requiring inferring characters' mental states (Pantelis, Byrge, Tyszka, Adolphs, & Kennedy, 2015; Richardson, Lisandrelli, Riobueno-Naylor, & Saxe, 2018), and playing interactive games that require reasoning about other's intentions (Chang, Smith, Dufwenberg, & Sanfey, 2011; Hampton, Bossaerts, & O'Doherty, 2008; Sul, Güroğlu, Crone, & Chang, 2017). These studies have identified a network of brain regions that appear to be reliably involved in mentalizing including the dorsomedial prefrontal cortex (dmPFC), temporal parietal junction (TPJ), superior temporal sulcus (STS), ventromedial prefrontal cortex (vmPFC), and posterior cingulate cortex (PCC) (Amodio & Frith, 2006; Blakemore, 2008; Frith & Frith, 2006; Van Overwalle, 2009).

While clever experiments have been instrumental in identifying the bounds of mentalizing, we still know very little about *how* these mental operations are specifically computed. A collection of empirical findings testing simple hypotheses are unlikely to reveal the complex processes underlying how people are able to infer the latent causes underlying others' behavior (Lewin, 1936; Newell, 2014). Instead, we might consider following trends in cognitive science and developing comprehensive and constrained models of cognitive architectures that are capable of predicting behavior across a range of different tasks (Newell, 1994; O'Reilly, Hazy, & Herd, 2016). Computational models provide a way to formally operationalize the mental processes that are hypothesized to underlie a specific cognitive operation (Jolly & Chang, 2019). This allows us to simulate behavior in various tasks and assess if the model can accurately account for how people behave in these environments. In addition, model representations of specific mental processes can be combined with neural recording techniques to identify the underlying neural circuitry, which can open up new avenues for improving our understanding of mentalizing (Cheong, Jolly, Sul, & Chang, 2017).

In this chapter, we will explore the burgeoning use of computational models to study mentalizing. The mathematical operationalization of these abilities provides a common language for researchers from disparate disciplines to bring together diverse perspectives (Jolly & Chang, 2019). In support of this interdisciplinary endeavor, we will briefly review some of the contributions to modeling mentalizing from the disciplines of cognitive science, computer science, economics, and cognitive neuroscience.

Game Theory of Mind

The ability to understand the minds of others has been modeled by economists over the past 70 years in the context of game theory (von Neumann & Morgenstern, 2007). Analogous to Lewin's proposition to model behavior as a function of the person and environment, game theory also models players and the structure of the game environment. Games refer to the mathematical descriptions of the strategies available to the players and of the payoffs resulting from those strategies. Additional details may include the sequence of play, the actions available to each player at each stage of the game, and the information available to each player. Players' beliefs are represented via probability distributions over actions, states, or other players' beliefs. Games involving two or more players and can be competitive or cooperative and played once (single shot) or repeatedly. One nice attribute of game theoretic modeling is that the behaviors of each player can be predicted without ever directly observing behavior. These predictions are referred to as solution concepts and often involve an equilibrium concept such as a Nash Equilibrium in which each player is assumed to know the strategy of the other players and no player can improve their payoff by changing their strategy (Nash, 1950). In this section, we will briefly discuss how the topics of strategic reasoning and psychological game theory have provided innovations in modeling mentalizing computations.

Strategic Reasoning

In economic games, players form strategies about which actions to take in the game. Players are assumed to be utility maximizers and will select the action that leads to the highest payoff. However, the payoffs in games are also determined by other players' actions. Thus, maximizing payoffs requires engaging in strategic reasoning, which may involve selecting an action based on first-order beliefs about the likelihood of other players taking a specific action. Some players may engage in even higher levels of strategic reasoning and form beliefs about other players' beliefs about their actions, which are referred to as second-order beliefs. Theoretically, there could be an infinite recursion of beliefs about beliefs. However, our brains do not possess infinite computational resources, and instead, there appear to be bounds on our cognitive capacity (Simon, 1956) and level of recursive reasoning or strategic sophistication (C. F. Camerer, Ho, & Chong, 2015; C. Camerer, Ho, & Chong, 2003; Stahl & Wilson, 1995; Yoshida, Dolan, & Friston, 2008).

The concept of bounded rationality in recursive mental state inference can be illustrated in the *beauty contest* game (Keynes, 1936; Nagel, 1998). In this game, players simultaneously choose a number between 0 and 100 and are informed that the guess closest to two-thirds of the average number wins a fixed prize. Players use a set of recursive thinking steps to choose the strategy expected to maximize payoffs, given their beliefs about the sophistication of other players. Zero-step players do not employ any strategic reasoning and choose their numbers randomly according to a uniform distribution. More sophisticated one-step players believe they are playing against zero-step players. The average of zero-step players' choices will be approximately 50, and thus will choose 33 (i.e., 2/3 of 50). Two-step players believe they are playing against a mixture of zero- and one-step players and reason about the proportions of each. In this game, payoff is maximized by adopting a strategy that is one level of sophistication higher than the average of all other players. If players can employ an infinite level of strategic sophistication, the Nash equilibrium in this game is to choose 0. In practice, behavior in the game can be approximated using a Poisson distribution with a $\lambda = 1.8$, which means that the distribution of sophistication levels is approximately zero, 17%; one, 30%; two, 27%; and three, 16% (C. Camerer et al., 2003). In other words, most players use a level 1 or 2 sophistication in their strategic reasoning. Using higher levels of sophistication in the beauty contest (Coricelli & Nagel, 2009) and other games (Yoshida, Seymour, Friston, & Dolan, 2010) has been shown to recruit increased activation in dmPFC when reasoning about other players, which is consistent with the interpretation that strategic reasoning involves representing others' mental states.

Psychological Game Theory

Modeling behavior in games does not only involve considering strategic reasoning, but also the goals and motivations of the players. Classic economic theory assumes that people receive utility solely from material payoffs. However, people have many other motivations that can influence behavior, such as reputational concerns and emotions (e.g., guilt, disappointment, regret, frustration, and anger), which are more abstract and difficult to quantify. Psychological Game Theory (PGT) is a framework to model motivations that depend on one's own or others' beliefs (Battigalli & Dufwenberg, 2009, 2019; Geanakoplos, Pearce, & Stacchetti, 1989). This framework has been useful in providing formal operationalizations of how emotions such as guilt (Battigalli & Dufwenberg, 2007; Chang et al., 2011; van Baar, Chang, & Sanfey, 2019) and anger (Battigalli, Dufwenberg, & Smith, 2015; Chang & Sanfey, 2013) can be represented as belief-dependent psychological payoffs in a utility function and can accurately predict behavior in a variety of cooperative games (Chang & Smith, 2015). For example, guilt can be modeled as a motivation to avoid disappointing a relationship partner by taking an action that is consistent with a player's second-order beliefs about what they believe their partner expects them to do. When playing games in the scanner, these belief-dependent emotions have been linked to regions of the brain that are involved in processing negative affect and error monitoring such as the anterior insula and anterior cingulate cortex (ACC) (Chang et al., 2011; Chang & Sanfey, 2013; van Baar et al., 2019).

In addition to modeling psychological payoffs associated with specific emotions, PGT has also been used to model how players perceive the intentions behind another player's actions. For example, Rabin (1993) proposed a model of intention-based reciprocity where kind intentions are reciprocated with kind actions, while unkind intentions are reciprocated with unkind actions. Though this tit-for-tat strategy had been previously described (Akerlof, 1982; Goranson & Berkowitz, 1966; Trivers, 1971), this was the first attempt to formally model the mentalizing process of inferring the hidden *intentions* of the other player from behavior in the game. This model was found to be a better explanation of behavior than the popular other-regarding preference model of inequity aversion (Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999) in a clever experimental design of mini-ultimatum games (Falk, Fehr, & Fischbacher, 2003). In this study, participants received a proposal of a 80–20% split of a pot of money and were asked to decide whether to accept the offer, in which they would receive the 20% portion, or reject it, in which case neither player would receive any money. The inequity-aversion model predicts that participants should reject the offer because that results in a more equitable outcome ($0–$0) than the 80–20 proposal. The experimenters provided an interesting twist by manipulating information about the counterfactual option that the proposer could have alternatively chosen. Most participants decided to reject the proposal when the 80–20 was chosen over a 50–50 split but chose to accept the *same* proposal when it was selected over a 90–10 split. Participants presumably inferred that the proposer had good intentions and accepting the offer would acknowledge and reciprocate this intention, while more selfish intentions of not selecting a 50–50 split would be punished. These types of other-regarding preferences are unlikely to be a stable preference that is inherent to a specific individual as they appear to evolve over the course of development. Children between the ages of 3–8 switch from being purely selfish to developing preferences for inequity aversion (Fehr, Bernhard, & Rockenbach, 2008). Preferences to consider others' intentions appear to shift later in adolescence, which is mediated by cortical maturation in the dmPFC (Sul et al., 2017).

Game theory has provided a unique opportunity to model mentalizing processes of representing others' beliefs, intentions, feelings, and desires in the context of constrained interactive economic games. Because these games are relatively simple, they are well suited for neuroimaging environments (Chang et al., 2011; Chang & Sanfey, 2013; Coricelli & Nagel, 2009; Hampton et al., 2008; King-Casas et al., 2005; McCabe & Smith, 2000; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; Sul et al., 2017; Yoshida et al., 2010). Furthermore, because these models represent different types of mentalizing operations, model-based fMRI analytic strategies (J. O'Doherty, Hampton, & Kim, 2007) have been fruitful in mapping these computations to brain activity in many regions known to be involved in representing others' mental states.

Machine Theory of Mind

How does a person come to intuitively know that a certain look from their partner means they had a tough day at work? Why can best friends have entire dialogues with each other without completing a single sentence? These impressive feats of mentalizing are a result of our impressive ability to continually *learn* statistical regularities from our experiences as we navigate the social world. In this section, we will explore how machine learning techniques have been used to model human learning about the beliefs, goals, and desires of others. One subdomain of machine learning that has shown promise in modeling artificial intelligence is reinforcement learning (RL). Similar to Lewin's model of behavior as a function of a person and environment, RL models agents acting within an abstracted environment. These types of models have witnessed extraordinary advances in recent years, merging supervised learning techniques such as deep neural network architectures to develop agents that outperform human experts in complicated strategic games, such as backgammon (Tesauro, 1994), chess (Campbell, Hoane, & Hsu, 2002), Atari (Mnih et al., 2015), go (Silver et al., 2018), and poker (Brown & Sandholm, 2019; Moravčík et al., 2017). While these models have achieved impressive performance that is beginning to generalize across tasks, they require considerable amounts of data and computational power, and it is unclear if the models are developing strategies that can help us understand how humans actually perform these computations. We will focus our discussion on some of the basic concepts from reinforcement learning to build an intuition for how these algorithms might help us model how people reason about others. We will provide examples in which people appear to use prediction errors to learn about others, their experiences, and optimal actions. We will also discuss how inverse reinforcement learning can be used to understand how we make predictions about another person's hidden mental state based on our observation of their actions.

Estimating Value

Formal reinforcement learning models were first inspired by learning experiments in psychology (Rescorla, Wagner, & Others, 1972) and were subsequently advanced by computer scientists (Sutton & Barto, 1998). RL models provide a computational approach to understanding and automating goal-directed learning and decision-making. In this framework, agents learn actions within an environment that maximize their overall cumulative reward. Predictions of future value V at timepoint $t + 1$ can be iteratively learned via prediction error, which is simply the difference between our experienced and expected rewards.

$$V_{(t+1)} \leftarrow V_{(t)} + \alpha \left[r - V_{(t)} \right]$$

The difference between the expected value V at time t and the experienced reward r constitutes an *error*, which can be reduced by taking a step toward the "target" (Sutton & Barto, 2018). We can think of the target as some true desired outcome presumed to come directly from the environment. An RL agent seeks an improved estimate of a signal's value by incrementally updating its old estimate, via a step size α, based on experienced errors. This provides a way to learn the value of a given signal in the environment via trial-and-error (Pearce & Hall, 1980; Rescorla et al., 1972). This simple prediction-error learning signal is reliably associated with activity in the ventral striatum when learning in simple tasks (Bartra, McGuire, & Kable, 2013; McClure, Berns, & Montague, 2003; J. P. O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003). Interestingly, this same learning system appears to aid in learning about an individual's moral character from repeated interactions in economic games such as trustworthiness (Chang, Doll, van't Wout, Frank, & Sanfey, 2010; Delgado, Frank, & Phelps, 2005; Fareri, Chang, & Delgado, 2012).

Observational Reinforcement Learning

RL provides a framework to not only learn about the moral character of another person, but also vicariously about the world via another person's experience. While observing an agent interact with their environment, prediction errors about the outcome can be computed for the agent rather than the observer. *Vicarious reinforcement learning* allows an observer to develop beliefs about the world without ever directly experiencing any outcomes (Golkar, Jangard, Tobler, & Olsson, 2019; Selbing & Olsson, 2017). These observed outcome prediction errors have been shown to be correlated with activity in the dorsal striatum and vmPFC (Burke, Tobler, Baddeley, & Schultz, 2010; Hill, Boorman, & Fried, 2016; Suzuki et al., 2012).

Sometimes, however, an observer can learn to *imitate* which actions to take, even when the outcomes or rewards for an observed agent's actions are not directly observable. Rather than learning the reward contingencies of the environment, imitation learning involves learning to take a particular action based on the extent to which another agent was observed to take that action in the past. Here, the value of a given action is computed through positive reinforcement if the action was performed by an observed agent, while unchosen actions are negatively reinforced. The difference between the action performed by an observed agent and the action that was expected by the observer constitute "action prediction errors" and can provide a learning signal similarly to observed reward prediction errors. In contrast to vicarious learning, imitation learning signals have been associated with a *different* set of brain regions including the dmPFC, dorsolateral prefrontal cortex (dlPFC) (Burke et al., 2010), and the inferior parietal lobule (Suzuki et al., 2012), a candidate location of mirror neurons in humans (Chong, Cunnington, Williams, Kanwisher, & Mattingley, 2008).

Inverse Reinforcement Learning

One potential drawback of both vicarious and imitation learning is the assumption that an observer possesses the same value function as the agent. However, there are many situations where two or more agents place vastly different values on the same thing. For instance, how do we learn about another person's food preferences after repeated dining experiences (Jern, Lucas, & Kemp, 2017)? To solve this, computational models of observational learning must permit inferences about another agent's goals independent of the observer's own preferences or value function. One way this could occur is by combining prior information with observed evidence of a given goal in order to update an observer's beliefs about an agent's goal in a Bayesian manner. Unlike standard RL, which attempts to learn the optimal actions given a reward function, *inverse reinforcement learning* attempts to recover the learned reward function for which an observed agent's actions would be optimal (Abbeel & Ng, 2004; Ng, Russell, & Others, 2000). This type of algorithm is particularly well suited to model hidden beliefs, goals, and desires from observing others' actions (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Jara-Ettinger, 2019; Jern et al., 2017; Koster-Hale & Saxe, 2013).

Inverse RL was recently examined in a study where observers were tasked with choosing between slot machines that would yield different types of food (Collette, Pauli, Bossaerts, & O'Doherty, 2017). Observers were only able to learn indirectly about the types of foods that were paid out through each of the slot machines by observing two other agent's actions, but not the outcomes of the agents' decisions. Importantly, observers knew that they shared food preferences with only one of the agents. The authors found that an inverse RL model explained participants' choices better than imitation learning. Consistent with previous studies (Burke et al., 2010), they found dmPFC activity to be involved in encoding the agent's (not the observer's) expected value of an action at the moment participants observed the agent chose that action. Because the agent's represented preferences differed from their own, the authors suggest that observers may have been simulating "what it would be like" to be the agent. Common mentalizing regions such as the TPJ and STS were also shown to track the degree to which observers updated their knowledge of the environmental reward structure. In addition to the mentalizing network, the dorsal striatum, lateral PFC, and pre-sensory-motor area tracked these updates on a trial-by-trial basis, suggesting that observational learning about reward distributions through inverse RL depends on regions commonly recruited for mentalizing, as well as areas generally recruited for experiential learning.

Bayesian Theory of Mind

A fundamental assumption made by game theoretic and RL models is that the ultimate goal of observed behavior is the maximization of rewards gained from one's environment. When an agent is faced with a decision about which action to take, they do so by intuiting a form of a *cost function* that, for each potential action in a given *state*, provides information to the agent of the resources (effort, time, money, etc.) required to perform it. The costs of taking an action are considered along with the agent's *reward function*, which maps environmental states to intrinsically beneficial rewards. The value of pursuing an action in a given state can be computed by the sum of all current and expected future rewards, discounted by the costs associated with taking the action. Optimizing this value function involves choosing actions that maximize this computation. This process of predicting an agent's future behavior on the basis of a set of rewards and constraints can be described as a Markov decision process (MDP).

The MDP framework captures the process by which agents use a model of their environment to decide which actions to take from its current state. As such, MDPs can represent a *causal* model for how an agent's actions change its environmental state on the basis of its subjective costs and rewards. For example, consider how a cleaning robot might navigate a room to pick up trash and avoid falling down the stairs (Fig. 1). The agent (cleaning robot) decides which actions to take (e.g., move right) in order to maximize rewards it gains from picking up trash items located at the goal state, while avoiding states that return negative values, such as falling down a flight of stairs in the room. The robot's battery is depleted (-1) by the energy required to move in any direction, so it must learn to optimize its behavior to efficiently reach its goals.

A criticism of using MDPs as a model of rational behavior, however, is that agents in the real world rarely possess full knowledge of their environment. For example, furniture or other items in the room may occlude the locations of trash rewards for the cleaning robot. Partially observable Markov decision processes (POMDPs) attempt to model the causal relationship between an agent's beliefs and their actions, given their uncertainty about the state of the world (Smallwood & Sondik, 1973; Sondik, 1978). Like MDPs, POMDPs attempt to find an optimal set of actions for an RL agent to maximize its reward. However, because the agent does

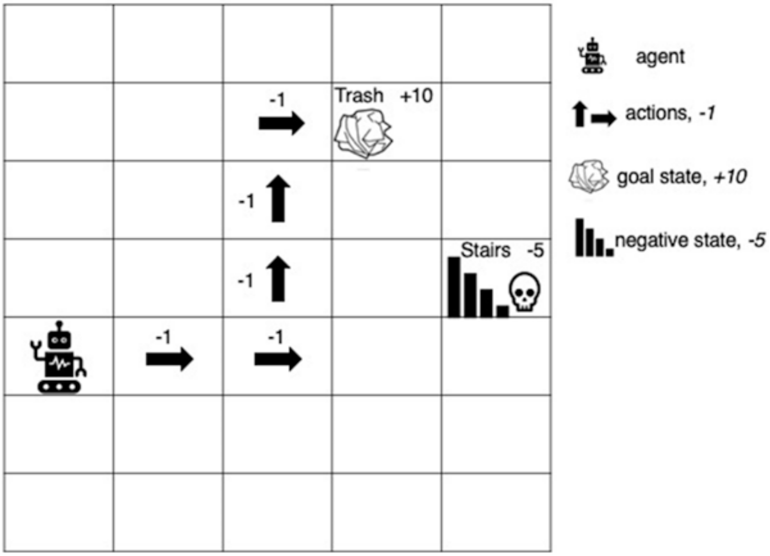


Fig. 1 Example of Markov decision process

not have full information about the world, the framework requires the ability to represent beliefs about the possible states of the world, which are updated based on the agent's observations. Thus, the expected future rewards from taking an action are based on the agent's internal beliefs about the state world.

POMDPs can provide a useful framework to model mentalizing. For example, Baker et al. (2017) use this framework to formulate a Bayesian theory-of-mind (TOM) model that attempts to infer an agent's likely goal, given its uncertain beliefs about the state of the world. In their experiment, they create a 2D gridworld environment of a parking lot containing two of three possible food trucks. A solid wall occludes the agent's view of the full environment so that it can only observe one truck from its starting state location. The authors represent the agent's observations in this spatial setting by indicating its visual line-of-sight, which allows the *observer* to know what information the agent possesses when making decisions. Participants are tasked with predicting the agent's beliefs and desires. The Bayesian TOM model attempts to capture an observer's hypothesis space of the agent's mind over the possible beliefs about which unseen food truck is on the other side of the wall, and the agent's possible reward function from simply observing the agent's observed actions. There are several components to the model. First, there is a term specifying the observer's prior belief $P(B_{t-1}, D, S)$ over the agent's initial beliefs B_{t-1} (which truck is beyond the wall), desires D (which truck the agent *hopes* is beyond the wall according to its preferences), and possible world state S (their position in the grid). Second, there is a likelihood function capturing what the observer believes the agent can see (percept P), given their position in the grid $P(P|S)$. Third, there is a likelihood function capturing the observer's model of how the agent updates their beliefs about which truck is behind the wall based on their percept, $P(B_t|P, B_{t-1})$. Finally, there is a likelihood function describing the observer's model of the agent's action plan A given their beliefs of which truck is behind the wall and desires $P(A|B_{\mathrm{t}}, D)$.

$$P\left(B,D,P,S,|\,A\right) \propto P\left(A\,|\,B_t,D\right) \times P\left(B_t\,|\,P,B_{t-1}\right) \times P\left(P\,|\,S\right) \times P\left(B_{t-1},D,S\right)$$

Across the space of possible beliefs and desires the agent could have, the observer evaluates the likelihood of generating the observed behavior given the hypothesized theory of the agent's mental state. Through a simple Bayesian update, the observer integrates this likelihood with the prior joint probability over mental states, yielding a posterior update that captures its inference of the agent's beliefs and desires (Baker et al., 2017; Baker, Saxe, & Tenenbaum, 2009). Overall, the authors find strong evidence that this model accurately captures experimental participants' judgments about the agent's beliefs and desires across a range of environments. Importantly, models that selectively lesioned representations of the agent's beliefs, desires, or percepts were unable to accurately capture observers' judgments, suggesting that people were jointly considering all of this information.

Summary

Building on Lewin's early social psychology theory, which attempted to describe behavior as a function of a person operating within a specific environment, the fields of game theory and reinforcement learning have made substantial advances in the past several decades developing formal models of people and environments. In this chapter, we have briefly reviewed some of the innovations within these domains to provide a general framework to model the goals of an agent (e.g., maximize reward, or minimize embarrassment), modules for how they might represent the mental states of other agents, such as their beliefs, goals, desires, intentions, and feelings, and modules for how to integrate internal goals and mentalizing computations to produce optimal policies to navigate the environment. The strength of this computational approach is that the framework and mathematical operationalization of these constructs facilitate collaborations across different laboratories and also scientific disciplines. Moreover, this framework is extensible, and new modules can be added to an agent or further refined to generalize across more complex environments (Jolly & Chang, 2019).

All of the models discussed in this chapter possess unique strengths and weaknesses and vary in their assumptions and explanatory power. It is important to note that although a given computational model may provide a good fit to behavioral data, this alone cannot definitively indicate if the model is accurately capturing the cognitive process employed by humans. Moreover, researchers may have diverging goals in modeling theory of mind. For some, the goal may be to develop systems that can accurately infer a person's mental state. Others may be more interested in characterizing how these computations are performed by humans and might be impacted by developmental and neuropsychiatric disease processes. The various classes of models introduced in this chapter may have varying utility in pursuing these different goals. For example, the Bayesian theory of mind and deep reinforcement learning models may be better suited to building industrial systems capable of inferring mental states based on observed behavior, language, and various types of sensing data, while the psychological game theory and simpler RL models may be more helpful in identifying how humans perform some of these computations. Each of these goals may also require different types of constraints. Theory of mind AI systems may be constrained by the available data and computational resources, while models of human mentalizing processes should be constrained by the structure of biological systems. These types of models can be useful in predicting systematic errors made by humans, or how these processes might be instantiated in the brain when combined with neuroimaging data. We note that mapping the models directly to neural activity is a challenging endeavor and almost always requires making additional assumptions that are difficult to validate. For instance, do the model predictions *linearly* map onto brain activity? Are model associations with *any* brain region equally informative, or do we have prior beliefs that some regions might be more likely to instantiate a process?

Though we believe computational approaches will aid in advancing our understanding of mentalizing, they are not without limitations, which are important to acknowledge. First, though there was early interest in integrating mathematical constructs into social psychological theory (Lewin, 1936, 1938), this vision never came into fruition as the field matured. Instead, the disciplines that have made substantial advances in this area such as economics, cognitive science, and computer science have emphasized technical sophistication and incorporated advanced mathematical and computing training in both undergraduate and graduate training programs. Psychology will have to make significant concerted efforts to update their training curriculum to be able to continue to contribute to this interdisciplinary endeavor (Jolly & Chang, 2019). Second, though the modeling is quite advanced, the mental operations and environments studied using this approach are necessarily oversimplified, which limits the sophistication of psychological inferences and generalizability of the findings. Most of the models discussed in this review are highly specific to a particular environment or game, and it is unclear how well they will generalize to new contexts. Third, as is becoming clear with the rapid advances afforded by the powerful function approximations made possible by deep neural network architectures, there is a tradeoff between predictive power and model interpretability. State-of-the-art models are now able to outperform the greatest strategic human minds at almost any type of complex game, but these models require enormous datasets, vast computational resources to train, and are not easily interpretable by humans. These deep learning approaches differ in their goals and as such they do not necessarily provide direct insight into the computations underlying *human* mental operations. However, we are optimistic that they may become increasingly more informative as they begin to share similar cognitive and computational constraints as humans. In summary, developing computational models of how humans perform mentalizing operations is an active and exciting area of research and much of its recent growth can be attributed to multidisciplinary efforts.

Acknowledgments This work was supported by funding from the National Institute of Mental Health R01MH116026 and R56MH080716 and the National Science Foundation CAREER 1848370.

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning* (p. 1). New York, NY: ACM.

Achim, A. M., Guitton, M., Jackson, P. L., Boutin, A., & Monetta, L. (2013). On what ground do we mentalize? Characteristics of current tasks and sources of information that contribute to mentalizing judgments. *Psychological Assessment, 25*(1), 117–126.

Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *The Quarterly Journal of Economics., 97*, 543. Retrieved from https://academic.oup.com/qje/article-abstract/97/4/543/1846076

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews. Neuroscience, 7*(4), 268–277.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour, 1*(4), 0064. s41562-017-0064 [pii].

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*(3), 329–349.

Baltaxe, C. A. (1991). Vocal communication of affect and its perception in three- to four-year-old children. *Perceptual and Motor Skills, 72*(3 Pt 2), 1187–1202.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*(1), 37–46.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage, 76*, 412–427.

Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *The American Economic Review, 97*(2), 170–176.

Battigalli, P., & Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory, 144*(1), 1–35.

Battigalli, P., & Dufwenberg, M. (2019). *Psychological game theory*. Retrieved from https://eller.arizona.edu/sites/default/files/Econ-WP-19-06.pdf

Battigalli, P., Dufwenberg, M., & Smith, A. (2015). *Frustration and anger in games*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2591839

Bennett, J. (1978). Some remarks about concepts. *The Behavioral and Brain Sciences, 1*(4), 557–560.

Blakemore, S.-J. (2008). The social brain in adolescence. *Nature Reviews. Neuroscience, 9*(4), 267–277.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *The American Economic Review, 90*(1), 166–193.

Brown, N., & Sandholm, T. (2019). Superhuman AI for multiplayer poker. *Science, 365*, 885. https://doi.org/10.1126/science.aay2400

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences of the United States of America, 107*(32), 14431–14436.

Call, J., & Tomasello, M. (1999). A nonverbal false belief task: The performance of children and great apes. *Child Development, 70*(2), 381–395.

Camerer, C., Ho, T., & Chong, K. (2003). Models of thinking, learning, and teaching in games. *The American Economic Review, 93*(2), 192–195.

Camerer, C. F., Ho, T.-H., & Chong, J. K. (2015). A psychological approach to strategic thinking in games. *Current Opinion in Behavioral Sciences, 3*, 157–162.

Campbell, M., Hoane, A. J., & Hsu, F.-H. (2002). Deep blue. *Artificial Intelligence, 134*(1), 57–83.

Chang, L. J., Doll, B. B., Van't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology, 61*(2), 87–105.

Chang, L. J., & Sanfey, A. G. (2013). Great expectations: Neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience, 8*(3), 277–284.

Chang, L. J., & Smith, A. (2015). Social emotions and psychological games. *Current Opinion in Behavioral Sciences, 5*, 133–140.

Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron, 70*(3), 560–572.

Cheong, J. H., Jolly, E., Sul, S., & Chang, L. J. (2017). Computational models in social neuroscience. In *Computational models of brain and behavior* (pp. 229–244). New York, NY: John Wiley & Sons. https://doi.org/10.1002/9781119159193.ch17

Chong, T. T.-J., Cunnington, R., Williams, M. A., Kanwisher, N., & Mattingley, J. B. (2008). fMRI adaptation reveals mirror neurons in human inferior parietal cortex. *Current Biology: CB, 18*(20), 1576–1580.

Collette, S., Pauli, W. M., Bossaerts, P., & O'Doherty, J. (2017). Neural computations underlying inverse reinforcement learning in the human brain. *eLife, 6*, e29718. https://doi.org/10.7554/eLife.29718

Coricelli, G., & Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 106*(23), 9163–9168.

Darwin, C. (1886). *The expression of the emotions in man and animals*. London: John Murray.

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience, 8*(11), 1611–1618.

Dennett, D. C. (1978). Beliefs about beliefs. *The Behavioral and Brain Sciences, 1*(4), 568.

Duchenne de Boulogne, G. B. A. (1876). *Mécanisme de la Physionomie Humaine ou Analyse Électro-Physiologique de l'Expression des Passions* (2nd ed.). Paris: Librairie J. B. Bailliere et Fils.

Falk, A., Fehr, E., & Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry, 41*(1), 20–26.

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience, 6*, 148.

Fehr, E., Bernhard, H., & Rockenbach, B. (2008). Egalitarianism in young children. *Nature, 454*(7208), 1079–1083.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics, 114*(3), 817–868.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50*(4), 531–534.

Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior, 1*(1), 60–79.

Goranson, R. E., & Berkowitz, L. (1966). Reciprocity and responsibility reactions to prior help. *Journal of Personality and Social Psychology, 3*(2), 227–232.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences of the United States of America, 105*(18), 6741–6746.

Harman, G. (1978). Studying the chimpanzee's theory of mind. *Behavioral and Brain Sciences, 1*, 576–577. https://doi.org/10.1017/s0140525x00076743

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology, 57*(2), 243–259.

Hill, M. R., Boorman, E. D., & Fried, I. (2016). Observational learning computations in neurons of the human anterior cingulate cortex. *Nature Communications, 7*, 12722.

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences, 29*, 105–110.

Jenkins, A. C., Macrae, C. N., & Mitchell, J. P. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proceedings of the National Academy of Sciences of the United States of America, 105*(11), 4507–4512.

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition, 168*, 46–64.

Jolly, E., & Chang, L. J. (2019). The flatland fallacy: Moving beyond low–dimensional thinking. *Topics in Cognitive Science, 11*, 433. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12404

Keynes, J. M. (1936). *The general theory of employment, interest and money*. Whitefish, MT: Kessinger Publishing.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science, 308*(5718), 78–83.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron, 79*(5), 836–848.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science, 354*(6308), 110–114.

Lewin, K. (1936). *Principles of topological psychology*. New York, NY: Martino Fine Books. https://doi.org/10.1037/10019-000

Lewin, K. (1938). *The conceptual representation and the measurement of psychological forces*. Durham, NC: Duke University Press. Retrieved from https://psycnet.apa.org/psycinfo/2008-10436-000/

Lindström, B., Golkar, A., Jangard, S., Tobler, P. N., & Olsson, A. (2019). Social threat learning transfers to decision making in humans. *Proceedings of the National Academy of Sciences of the United States of America, 116*, 4732. https://doi.org/10.1073/pnas.1810180116

Martin, J., Rychlowska, M., Wood, A., & Niedenthal, P. (2017). Smiles as multipurpose social signals. *Trends in Cognitive Sciences, 21*(11), 864–877.

McCabe, K. A., & Smith, V. L. (2000). A comparison of naïve and sophisticated subject behavior with game theoretic predictions. *Proceedings of the National Academy of Sciences, 97*(7), 3777–3781.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron, 38*(2), 339–346.

Mitchell, J. P., Heatherton, T. F., & Macrae, C. N. (2002). Distinct neural systems subserve person and object knowledge. *Proceedings of the National Academy of Sciences of the United States of America, 99*(23), 15238–15243.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533.

Moravčík, M., Schmid, M., Burch, N., Lisý, V., Morrill, D., Bard, N., … Bowling, M. (2017). DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science, 356*(6337), 508–513.

Müri, R. M. (2016). Cortical control of facial expression. *The Journal of Comparative Neurology, 524*(8), 1578–1585.

Nagel, R. (1998). A survey on experimental 'beauty contest games': Bounded rationality and learning. In D. Budescu, I. Erev, & R. Zwick (Eds.), *Games and human behavior: Essays in honor of Amnon Rapoport*. Mahwah, NJ: Lawrence Erlbaum Associates.

Nash, J. F. (1950). Equilibrium points in N-Person games. *Proceedings of the National Academy of Sciences of the United States of America, 36*(1), 48–49.

Newell, A. (1994). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Newell, A. (2014). *You can't play 20 questions with nature and win: Projective comments on the papers of this symposium*. Retrieved from https://kilthub.cmu.edu/articles/You_can_t_play_20_questions_with_nature_and_win_projective_comments_on_the_papers_of_this_symposium/6612977/files/12105638.pdf

Ng, A. Y., Russell, S. J., & Others. (2000). Algorithms for inverse reinforcement learning. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers Inc.

O'Doherty, J., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences, 1104*(1), 35–53.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron, 38*(2), 329–337.

O'Reilly, R. C., Hazy, T. E., & Herd, S. A. (2016). The Leabra cognitive architecture: How to play 20 principles with nature. In *The Oxford handbook of cognitive science* (p. 91). Oxford: Oxford University Press.

Pantelis, P. C., Byrge, L., Tyszka, J. M., Adolphs, R., & Kennedy, D. P. (2015). A specific hypoactivation of right temporo-parietal junction/posterior superior temporal sulcus in response to socially awkward situations in autism. *Social Cognitive and Affective Neuroscience, 10*(10), 1348–1356.

Parkinson, C., Walker, T. T., Memmi, S., & Wheatley, T. (2017). Emotions are understood from biological motion across remote cultures. *Emotion, 17*(3), 459–477.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87*(6), 532–552.

Pylyshyn, Z. W. (1978). When is attribution of beliefs justified? *Behavioral and Brain Sciences, 1*, 592–593. https://doi.org/10.1017/s0140525x00076895

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review, 83*(5), 1281–1302.

Rescorla, R. A., Wagner, A. R., & Others. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications, 9*(1), 1027.

Rosenthal, I. A., Hutcherson, C. A., Adolphs, R., & Stanley, D. A. (2019). Deconstructing theory-of-mind impairment in high-functioning adults with autism. *Current Biology, 29*, 513–519.e6. https://doi.org/10.1016/j.cub.2018.12.039

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science, 300*(5626), 1755–1758.

Saxe, R., & Kanwisher, N. (2003/2008). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *NeuroImage, 19*(4), 1835–1842.

Selbing, I., & Olsson, A. (2017). Beliefs about others' abilities alter learning from observation. *Scientific Reports, 7*, 16173. Retrieved from https://www.nature.com/articles/s41598-017-16307-3

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., … Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science, 362*(6419), 1140–1144.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129–138.

Smallwood, R. D., & Sondik, E. J. (1973). The optimal control of partially observable Markov processes over a finite horizon. *Operations Research, 21*(5), 1071–1088.

Sondik, E. J. (1978). The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. *Operations Research, 26*, 282–304. https://doi.org/10.1287/opre.26.2.282

Stahl, D., & Wilson, P. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior, 10*(1), 218–254.

Sul, S., Güroğlu, B., Crone, E. A., & Chang, L. J. (2017). Medial prefrontal cortical thinning mediates shifts in other-regarding preferences during adolescence. *Scientific Reports, 7*(1), 8510.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 135). Cambridge, MA: MIT Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., … Nakahara, H. (2012). Learning to simulate others' decisions. *Neuron, 74*(6), 1125–1137.

Tesauro, G. (1994). TD-Gammon, a self-teaching Backgammon program, achieves master-level play. *Neural Computation, 6*(2), 215–219.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35–57.

van Baar, J. M., Chang, L. J., & Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature Communications, 10*(1), 1483.

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30*(3), 829–858.

von Neumann, J., & Morgenstern, O. (2007). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655–684.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103–128.

Yoshida, W., Dolan, R. J., & Friston, K. J. (2008). Game theory of mind. *PLoS Computational Biology, 4*(12), e1000254.

Yoshida, W., Seymour, B., Friston, K. J., & Dolan, R. J. (2010). Neural mechanisms of belief inference during cooperative games. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 30*(32), 10744–10751.

From Neurons to Knowing: Implications of Theoretical Approaches for Conceptualizing and Studying the Neural Bases of Social Understanding

Jeremy I. M. Carpendale, Ulrich Müller, Charlie Lewis, and Beau Wallbridge

The viability of approaches taken to studying the neural bases of social understanding, also referred to as "mentalizing," depends on how social understanding is conceptualized, which, in turn, depends on researchers' underlying sets of philosophical assumptions, or worldview. In order to trace the implications for studying the neural bases of "mentalizing," we will characterize two families of approaches to social understanding and outline the two worldviews on which they are based. The first, which we critically assess, has been referred to as cognitivist, mechanistic, individualistic, or Cartesian worldview. As an alternative, we introduce the process-relational approach that we endorse.

Ideas from developmental science and cognitive science tend to be incorporated without careful evaluation into social cognitive neuroscience research. We advocate a more critical assessment of such ideas. We provide such an analysis with the examination of two commonly made assumptions: (1) that some forms of knowledge are innate and (2) that thinking is computation. These assumptions are not based on evidence from neuroscience; instead, they are drawn from perspectives in developmental psychology. These are theoretical claims based on a particular interpretation of data; the data do not provide the theory; rather, assumptions color the way data are collected and interpreted. The data are theory-laden, situated within a theoretical framework (Hanson, 1958). In particular, social cognitive neuroscience

J. I. M. Carpendale (✉) · B. Wallbridge
Department of Psychology, Simon Fraser University, Burnaby, BC, Canada
e-mail: jcarpend@sfu.ca; bwallbri@sfu.ca

U. Müller
Department of Psychology, University of Victoria, Victoria, BC, Canada
e-mail: umueller@uvic.ca

C. Lewis
Department of Psychology, Lancaster University, Lancaster, UK
e-mail: c.lewis@lancaster.ac.uk

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_16

tends to be assimilated to the information processing and computational view of the mind, regardless of whether this fits with the neuroscience.

Although neural activity is clearly necessary for social understanding, beyond mapping the activity of brain regions, learning about social understanding will require taking the step described in our title “from neurons to knowing” because social understanding does not exist at the causal level of the transformation of energy and the firing of neurons in response to stimuli. Instead, it emerges within human experience as a form of knowledge that is learned in engagement with the world and other people. We return to this issue at the end of the chapter. First, we explicate the sets of assumptions underlying two families of theories (for somewhat similar approaches, see Di Paolo & De Jaegher, 2012; Fuchs, 2011; Kiverstein & Miller, 2015; Sameen, Thompson, & Carpendale, 2013).

Tracing the Implications of Philosophical Assumptions: Two Worldviews

Presuppositions matter for research and the interpretation of results. This can be seen in the case of social cognitive neuroscience in the way that views from developmental psychology and cognitive science are adopted, often uncritically. We briefly outline two sets of preconceptions or worldviews and trace out their implications for conceptualizing and studying the neural bases of social understanding. We critique the first worldview and, as an alternative, introduce a second approach. The first approach we critically examine can be referred to as cognitivist, mechanistic, or individualistic because it begins by assuming the individual mind as the starting point that is thought to be private and accessible only to the self. It has also been referred to as Cartesian because this view of the mind was famously articulated by Descartes, although it has a long history and can already be found in the earlier writings of Saint Augustine. This approach assumes a split between a preexisting mind and the world, so it is also referred to as a split approach (Overton, 2015). From this perspective, it is assumed that infants have experience of their own separate individual minds, and thus in developing social knowledge, they face what is referred to as the “problem of other minds.” Accordingly, they must learn that others have minds just like themselves. This is claimed to be a difficult task because of the Cartesian assumption that action is caused by invisible mental states such as beliefs, desires, and intentions that underlie and cause outer physical behavior, yet somehow children must learn about such invisible entities (e.g., German & Leslie, 2004).

Given these preconceptions, several apparently quite different solutions are already built into this assumed starting point. First, children could formulate a theory about the world and other people, as claimed in the “theory” theory (e.g., Gopnik & Wellman, 2012), or they could be born with such a theory or knowledge, as assumed in innate module theories (e.g., German & Leslie, 2004). There are combinations of these approaches as in the claim that infants are born with a starting point theory that they then revise (Gopnik & Wellman, 2012). Another proposed solution is that since it is assumed that children start off with their own mind, they could apply their own inner experience to understanding others, referred to as the simulation approach or reasoning by analogy (e.g., Harris, 1991; Meltzoff, 2007). The first solution, "theory" theory, begins with the assumption that children can form theories or are born with a theory as a starting point that is then modified. This relies on the idea of innate knowledge, which we examine below, as does the innate module approach. The latter solution, the simulation approach, also assumes that infants begin with a mind and can use their own experience to model what the other person is feeling or thinking. The basic idea is that when the infant sees a bodily expression of another person, she can infer, based on the similarity between this expression and her own bodily expression, that the other person must feel what the infant is feeling when enacting the bodily expression (e.g., Gordon, 1986; Meltzoff, 2007; Meltzoff, Gopnik, & Repacholi, 1999). However, as has long been recognized (Scheler, 1954), the simulation theory presupposes what it aims to explain. Zahavi (2008, p. 517) succinctly summarizes the main flaw identified by Scheler:

> In order for the argument to work, there has to be a similarity between the way in which my own body is given to me, and the way in which the body of the other is given to me. But if I am to see a similarity between, say, my laughing or crying and the laughing or crying of somebody else, I need to understand the bodily gestures and behavior as expressive phenomena, as manifestations of joy or pain, and not simply as physical movements. If such an understanding is required for the argument of analogy to proceed, however, the argument presupposes that which it is supposed to establish. To put it differently, in some cases we do employ analogical lines of reasoning, but we only do so when we are already convinced that we are facing minded creatures but are simply unsure about precisely how we are to interpret the expressive phenomena in question.

Furthermore, for simulation theory to work, infants must already be equipped with rather sophisticated cognitive abilities, including the ability to reflect on themselves. One necessary condition for self-reflection, however, is that social interaction with others has attained a level of complexity that allows one to take the perspective of the other on the self (Mead, 1934; Müller & Carpendale, 2004), so the ability to reflect on experience is better viewed as a developmental achievement or outcome of social interaction, rather than its starting point (Baldwin, 1906). Alternatively, simulation could be viewed as some kind of primitive, reflexive mirroring mechanism, but if this position is taken, then simulation can help little to explain how children come to understand the meaning of the others' bodily expressions (for further discussion and review, see, e.g., Carpendale & Lewis, 2015; Zahavi, 2008).

Although these theories might seem quite different, they are actually all attempts at solutions to the same problem resulting from philosophical preconceptions. Thus, the difficulty we point out is tied to the way the problem is described in the first place. These are philosophical assumptions, and there is nothing empirical about how they are arrived at, but they do have important empirical implications for research and the interpretation of results. These worldviews cannot be tested with critical experiments, but they still can be critically evaluated (e.g., Jopling, 1993;

Overton, 2015). Worldviews cannot be directly falsified by empirical tests; the relation between the philosophical assumptions underlying worldviews and data is mediated by a number of different assumptions (e.g., methodological, statistical, etc. assumptions) that all can be brought to bear when empirical evidence is inconsistent with theoretical predictions and needs to be discounted. What is needed is to question these initial assumptions (Jopling, 1993), rather than solve the problem as it is presented.

The alternative worldview that we review, which has come to be known as process-relational, also has a long history with roots in Aristotle and the writings of scholars such as James Mark Baldwin, John Dewey, George Herbert Mead, Jean Piaget, Charles Sanders Peirce, Lev Vygotsky, and Heinz Werner, and Ludwig Wittgenstein's later work (Bernstein, 2010). More recently, analyses have been provided by Willis Overton and Richard Lerner, among others (e.g., Lerner, Agans, DeSouza, & Hershberg, 2014; Overton, 2015). From this perspective, the beginning point is a process of interaction, not a preexisting mind. An initial question is how does this process unfold? We will attempt to answer this question below.

Worldviews have implications at all levels from genes to justice and proteins to politics. Here we focus on the implications for social cognitive neuroscience. The next step is to move from these broad philosophical assumptions to how this matters for research on the neural bases of social understanding. We will take two relevant issues as examples: (1) claims that some forms of knowledge are innate and (2) the assumption that thinking is computation, based on the information processing approach.

Can Knowledge Be Innate?

Claims that forms of thinking are "specified by our genetic program" were more common 20 years ago (e.g., Pinker, 1997, p. 21) than they are today. Currently, such claims tend to be made more ambiguously and less explicitly with statements that thinking is "hard-wired," or has "biological foundations," or that infants are "endowed with" or "equipped with" forms of knowledge, implying that infants are born with forms of knowledge. These are still claims that forms of thinking are innate. Yet the explicit claim is still often made, even recently. For example, as their third question concerning what they think neuroscience can reveal about cognition and its sources, Sherry and Saxe (2016, p. 322) consider "what is the initial state of the human mind (i.e., what is specified genetically?)." Given the perpetuation of points like this, it is important to examine its implications, since it reflects a perspective adopted in neuroscience research and also in discussions in developmental psychology and cognitive science.

The claim that the initial state of the human mind is innately specified is not backed up by explanations regarding exactly how it is possible to show a link from molecules to minds. Thus, what is needed in order to evaluate such claims is some rudimentary understanding of genetics (Meaney, 2010). Work in biology has moved on in the last 50 years from the view that brain maturation is prespecified. Current thinking in biology is that neural interconnectivity is shaped through experience (e.g., Stiles, 2009; Stiles, Brown, Haist, & Jernigan, 2015). Furthermore, it has been argued that the term innate is ill-defined and has many different uses, all of which are problematic (see Mameli & Bateson, 2006). Genes are one factor in complex developmental systems. Particular genes can have varying effects depending on what other factors are present in the cell—in some cases, these effects can vary as widely as from cell life to cell death (Meaney, 2010). Fisher (2006, p. 270) argues that to begin to understand the complex processes in getting from genes to thinking we have to understand that "genes do not specify behaviours or cognitive processes." Instead, genes are involved in producing "regulatory factors, signaling molecules, receptors, enzymes and so on that interact in highly complex networks, modulated by environmental influences" (Fisher, 2006, p. 270). In fact, epigenetics is the study of the many factors influencing how genes are expressed, including social experience.

At birth, the brain is structured in terms of differences between regions such as in density and types of neurotransmitters (Stiles, 2009; Stiles et al., 2015). Patterns of neural interconnectivity, however, are shaped by experience, and the incredibly vast number of synapses in the human brain increases and decreases over development and cannot be prespecified. An influential approach that takes into account the complex nature of brain development is neuroconstructivism, which emphasizes the role of experience in brain development (Mareschal et al., 2007). From this perspective, skills do not preexist in particular brain regions nor do they typically develop anywhere in the brain. Instead, the differences between brain regions result in some of regions being particularly well suited for responding to particular experience, so that, for example, certain regions tend to develop language, unless they are damaged, in which case skills such as language may develop in other regions (Bates, 1999, 2005).

Most contemporary researchers assume that biological and social factors interact. However, there is still a crucial difference between different ways of conceptualizing this interaction. Interaction can be thought of as occurring between two preexisting entities such as social and biological factors or, as in the classic distinction, between nature and nurture. From this perspective, it is assumed that it should be possible to determine the relative contribution of genes and environment to development. However, Meaney (2010) emphasizes that genes and environment cannot be separated meaningfully: "Attempts to parse the influence of genomic and environmental influences on the expression of complex traits are inconsistent with even the most rudimentary understanding of gene function" (Meaney, 2010, p. 69, see also Gottlieb, 2007). A second, more thorough-going view of interaction is that it is actually not possible to separate preexisting factors that interact. Instead, when we look closely at any aspect of development, we see how thoroughly interwoven the dimensions are and how biological and social factors mutually create each other. It is interaction that is primary.

The view that interaction is primary is essential to Developmental Systems Theory (DST), which eschews the dichotomy between nature and nurture. Instead, the position is that these factors can only be artificially abstracted out of a thoroughly integrated matrix. In the abstract, we can talk about social and biological factors, but when we look at actual concrete examples, it becomes impossible to clearly distinguish them (Gottlieb, 2007; Jablonka & Lamb, 2014; Lewontin, 1983/2001; Lickliter & Honeycutt, 2015; Lickliter & Witherington, 2017; Oyama, Griffiths, & Gray, 2001).

There are deep roots to the intuition that something is fixed in development and that this is attributed to something at the biological level. But regularity in outcome can be the result of consistency in a complex developmental system rather than to fixed preexisting information, just as a mature forest is a regular outcome given certain climates and combinations of species, yet it is the result of a system of interacting factors rather than preexisting information. Thus, to study social understanding, we have to focus on the process of development.[1]

Is Thinking Computation?

The second issue we examine is the claim that thinking is computation. It is assumed that it does not matter if computation is performed by a computer or a brain. This is sometimes combined with the previous claim of innate knowledge. For example, "the mind is a system of organs of computation, designed by natural selection to solve the kinds of problems our ancestors faced in their foraging way of life" (Pinker, 1997, p. 21), or "we inhabit mental worlds populated by the computational outputs of battalions of evolved, specialized neural automata" (Tooby & Cosmides, 1995, p. xi). Onishi and Baillargeon (2005, p. 257) "assume that children are born with an abstract computational system that guides their interpretation of other's behavior." More recently, Sherry and Saxe (2016, p. 322) argue that neuroscience can make a "deep contribution to cognitive science, as it provides constraints on the algorithms by which information is transformed during processing and inference." They go on to questions "regarding which representations and computations are present innately and which are constructed from specific kinds of experiences"

[1] Because we discuss regularity in developmental outcomes, does this mean these traits are "innate"? No. Such regularity can be due to the outcome of processes of interaction within developmental systems. Regularity in outcomes is actually the third of the 26 definitions of "innate" (all problematic) considered by Mameli and Bateson (2006). They point out that regularity in outcomes is consistent with the trait being learned, not innate. We have argued against a separation between genetic and environmental factors. Instead, there are a host of biological factors (in addition to DNA) that mutually create and bi-directionally interact with environmental factors. We object to the word innate because it is ill defined, and it is often viewed as providing an explanation. Yet, as Spencer et al. (2009, p. 80, italics in original) point out, any predisposition claimed to be innate, "like any other characteristic of an animal, must *develop*, and it is important to study the process through which this occurs" (see also Lehrman, 1953). These predispositions may develop prior to birth or after birth, but they do develop (Spencer et al., 2009). Claims of innateness are not explanations; instead, they are placeholders or promises for developmental explanations.

(p. 336). Again, this presupposes that thinking is computation. These assumptions are stated but not defended from the criticism mentioned above.

These sorts of claims require a way of moving from the level of genes to patterns of neural interconnectivity that are conceptualized in terms of computation, which we have questioned in the previous section. Beyond this problem, a number of criticisms have been leveled against the computational theory of mind (e.g., Heil, 1981, 1998; Shanker, 1998). The issue that we will focus on is its failure to account adequately for meaning. Thinking is about the world and therefore is meaningful. If computation is thought of as the manipulation of symbols that are meaningfully linked to the world, then how do such symbols acquire meaning? This is known as the symbol-grounding problem. In the case of computer programs, the person using the program assigns meaning to the symbols, but this cannot be the case if human thinking is conceptualized as computation because there cannot be a small person (a homunculus) in the system assigning meaning (e.g., Heil, 1998; Kenny, 1971/1991). Such a homunculus account would just defer the explanation rather than provide one.

Computation is a mechanical process, and the meaning of symbols must be fixed in what are referred to as "computer languages." But, although the same word is used, computer languages are not at all similar to any natural language used by humans. Meaning in human communication and thinking is not fixed. This is the way human languages work. Wittgenstein (1967) used a series of examples to show that meaning is not fixed to representations. For example, he suggests that the reader consider a picture of a boxer, which might be taken as a classic example of a representation, akin to an image that comes to mind. This picture could mean the correct way to stand, or how not to stand, or it could refer to a particular boxer or, indeed, any number of other possible meanings. The same problem applies to any representation from words to sentences or pictures because they can all have multiple meanings and can be used to convey many different messages when set within different social relations. This problem applies to mental representation as well because this does not avoid this issue that meaning is not fixed (Goldberg, 1991; McDonough, 1989, 1999).

One suggestion is that mental representations acquire meaning through being causally linked to the world (e.g., Perner, 1991). That is, if we open our eyes and form a mental representation of an oak, for example, that would be meaningful because it is caused by the world, by the actual tree. Does this work to provide meaning? There are a number of reasons why this is problematic. For example, a camera can record an image of the oak, and it could be said to store information in one sense of that word and the information is caused by the world. But the camera doesn't know anything about trees or anything else, so there is something missing in explaining knowledge in this way.

Furthermore, Putnam (1988) examines this claim that mental representations acquire meaning through a causal connection with the world with a thought experiment. He suggests that we imagine a person seeing a tree and forming a mental representation of it. Then imagine a photograph of this tree being dropped on a planet on which no trees grow. A person picking up and looking at that image of the tree would form a mental representation of it and perhaps could be said to have some sort of knowledge of a tree because there is a causal chain between the tree, the photograph, and the mental representation. But now take it one step further and break that causal chain. Suppose instead that the image dropped on the planet was the result of accidently spilled paints that just happened to look exactly like the tree. In this case, the person looking at the image would have the same mental representation as the person looking at the photograph, yet would have no knowledge of trees. It is for this reason that such representations cannot explain knowledge, and why the causal theory of representation breaks down.

The computational view of the mind is one application of a more general perspective known as information processing approach. It might be possible to endorse an information processing approach without accepting the computational view. The basic idea underlying information processing approaches is that thinking can be conceptualized as the processing of information. Within this framework, the brain is assumed to process "input" and lead to "output." However, both the terms information and processing have quite different meanings in the context of people compared to computers. The information processing approach overlooks and conflates those meanings. We must develop and explain them in order to evaluate the approach.

Processing information when talking about people refers to understanding, evaluating, judging, deciding, and so on, and often finally acting. But none of this applies in the case of computers. Processing in that case refers to the transformation of series of digits through an algorithm designed by a programmer. This is a mechanical procedure like a slide rule or an abacus. The computer doesn't understand any of it at any point. It might be argued that computers can act if they are linked to some output device. But the production of movement is not the same as action. An automatic door opener is not really very polite, and if it doesn't work this doesn't mean that it is rude, but just broken.

Another approach to the central issue of meaning is through considering how information is conceptualized. It might seem obvious that we perceive the world and thus acquire information and that we process that information in various ways. A computer can be described as doing the same thing. But, in fact, there are important distinctions to be made. An airline schedule contains information and a camera can record and store information, but this a different sense of the term compared to saying that a person has information in the sense of knowing something. The problem is the relation between these two senses. To conflate the two meanings is to conceal what has to be explained (Kenny, 1971/1991). Light can enter a person's eye and cause various reactions and transformations in the retina and optic nerve leading to the brain, and resulting finally in seeing. But this outcome is the problem requiring an explanation.

Certainly, we are informed through our experience. But conceptualizing information as preexisting already assumes a theory of knowledge. It is the view of knowledge that John Dewey (1960) labeled a "spectator theory of knowledge" and Piaget (1970) referred to as a "copy theory." The assumption is that knowledge is passively received from the world through an individual's senses in the way that a camera records information. A problem is that, although we could think of the camera as storing information, this is quite different from a person seeing something because a camera does not know or understand anything and a person does. It is this understanding that has to be explained. There is no way to check such copies of reality except by forming another copy, so this still doesn't solve the problem of getting at reality (Chapman, 1999; Piaget, 1970). An alternative view of knowledge from the process-relational perspective is known as constructivism, which we introduce below.

Hobson (2002, p. xiv) points out that "computers don't understand anything nor do they care." They are passive machines not linked to the world in a way that they can form a meaningful connection. What we will focus on is the problem of how the world becomes meaningful for the child, and this leads to the positive account we will turn to next.

A Process-Relational Account

We have argued that the neural interconnectivity involved in social understanding cannot be completely prespecified by genes and instead develops through interactive experience with others and the world. Furthermore, what develops is not an "abstract computational system." So, how should the neural basis of social understanding be conceptualized?

It follows from the arguments in the previous sections that we need an adequate account of meaning and how it develops—how the world becomes meaningful for the child—in order to explain human intelligence. For this, a theory of knowledge is needed, which includes an understanding of the links between the child and world. From the perspective of a process-relational worldview, knowledge is constructed through interaction. According to the view of knowledge known as constructivism, the beginning point is with the infant's actions set in the physical and social world. The focus is on relations and process rather than a preexisting mind. Infants are actively immersed in relations with the world. They interact with the world and gradually learn to anticipate what will happen as a result of their actions. They develop patterns of activity, or schemes, to do with aspects of the world such as objects. They learn to coordinate their senses such as vision with actions such as grasping. Through this process, they learn the interactive potential of the world they experience. They learn what they can do with objects such as grasp them and suck them or drop them and so on. Through such experience, objects acquire significance or meaning for the infants. They perceive them in terms of what they could potentially do with them. This is sensorimotor knowledge described by Piaget (1936/1952, 1970). This process typically occurs in a thoroughly social context because human infants are born relatively helpless and they must be cared for, thus guaranteeing a social environment in which infants develop (Portmann, 1944/1990).

From a process-relational perspective, it is not assumed that infants are born with a mind, but that skills in thinking develop within social relations. Therefore, the question becomes what factors start the social process going within a human developmental system. Rather than positing an “abstract computational system” in order to explain the development and evolution of social intelligence, we examine how different aspects of an evolved developmental system enable the development of social intelligence. One aspect is the potential for the nervous system to be shaped by the sort of social interaction human infants experience. A second aspect is that infants inherit not only their genes but also their physical and social environments (Jablonka & Lamb, 2014), within which others are readily attentive and responsive to infants’ activities. The third aspect to consider is the adaptations that result in children’s social experience. That is, it is crucial to think about adaptations that result in social experience of particular kinds. For example, human infants are born early and relatively helpless, which results in a necessarily social environment in which they develop within social relations. Furthermore, even newborn infants are interested in human eyes, and the tendency to look at their caregivers’ eyes directs them to important sources of social information (Farroni, Massaccesi, Pividori, & Johnson, 2004). In addition, human eyes are highly salient compared to other primates because the dark pupil is surrounded by a white area (Kobayashi & Kohshima, 2001), thus supporting social engagement (Senju & Johnson, 2009; Tomasello, Hare, Lehmann, & Call, 2007). These are examples of adaptations that result in the social experience within which human infants develop. Infants’ sensitivities draw them toward as well as elicit aspects of the environment within which they develop further skills. These new skills, in turn, result in new experiences, in a constantly changing bidirectional process (Carpendale, Frayn, & Kurcharczyk, 2017).[2]

Infants learn to coordinate their actions with others and ascertain how others respond to their actions. From this perspective, communication is viewed as developing through the emergence of shared patterns of interaction within which infants and caregivers gradually learn to coordinate their interactions (Clark, 1978; Mead, 1934). For example, typically developing infants get better at coordinating their actions with those of their caregivers. By 2–4 months, they learn to stiffen their bodies in anticipation of being picked up by their caregivers (Reddy, Markova, & Wallot, 2013). Interactions such as these form the context within which infants then come to anticipate what happens next, including how others respond to their actions. It is within these shared patterns of interaction that more complex intentional communication develops (Canfield, 2007). As an example of the development of requests, infants experience their caregivers as sources of comfort when they are distressed. Then, as they learn to coordinate their reaching action, they can extend their arms toward their caregivers if they are distressed and desire comfort through being held by them. At the beginning, this is not an intentional act of

[2] Because we discuss the newborn infant’s embodiment and action tendencies, does this mean that we are somewhat nativist, and that nativists just specify richer innate structures? No. Nativists claim that infants are born with innate knowledge. For us, infants are born after 9 months of development with abilities, sensitivities, action tendencies, ways of being embodied within a necessarily social world due to their helplessness, and so on. This sets up the human developmental system in which the development of communication, language, and thinking occurs. The biology creates the social, which shapes the biology in a bidirectional process over time.

communication, but their desire is manifest in their action, and thus it does function to communicate this desire to their caregivers. Caregivers typically respond by picking up their infant, who then learns to anticipate this response and to grasp the meaning their action has for others. Within such a process, infants can begin to communicate intentionally with the expectation of the response of being picked up (Mead, 1934). This description of what is known as the "arms up" gesture is a common and early developing gesture used to make requests (Service, Lock, & Chandler, 1989) and is an example of how infants can elicit aspects of their environments, in this case contingent responsiveness from their caregivers, that facilitate further development. Similarly, infants learn to make requests for objects after they have learned to coordinate their reaching actions and learn what typically happens when these actions are performed in the presence of others (e.g., Carpendale & Ten Eycke, 2020). A crucial part of this account is that infants and caregivers enjoy interacting with each other—interactions are infused with emotions. Mutual joy forms the basis around which infants can learn to share attention on objects or events in enjoyable ways with others (Bates, Camaioni, & Volterra, 1975). These interactions then form the basis upon which further social and communicative development takes place (Rodríguez, Moreno-Núñez, Basilio, & Sosa, 2015). What infants develop is a lived sensorimotor, practical understanding of interacting with others and becoming better at both anticipating others' responses and eliciting desired responses from others.

Meaning is conveyed by relying on shared social relations (Winch, 1958) within which infants learn to use gestures such as the arms up gesture, and more complex gestures such as pointing (Carpendale & Carpendale, 2010). As infants learn to use gestures to intentionally communicate within shared patterns of interaction, they can begin to integrate words into these shared routines to communicate what was previously communicated through gestures. For example, infants could begin to use a word like "want" with requesting, and "look" and "see" within situations of sharing attention (Carpendale & Ten Eycke, 2020). Notice that these are mental state terms and their meaning is based on social relations. More complex mental state terms such as think, know, forget, and decide are based on more complex social relations. For example, a 3-year-old can learn to use the word "forgot" in the context of her mother bringing her toast for breakfast but without the expected jam. From a process-relational perspective, a child's use of "forgot" in this context refers to the readily observable pattern of activity based on expectations of what typically happens, and the word can be learned this way.

This approach to how children develop social understanding and learn to use mental state terms contrasts sharply with many "theory of mind" approaches that assume a Cartesian view of the mind according to which mental state terms are assumed to refer to causal inner mental entities that are separate from action, and learning the word requires mapping new words onto such inner mental entities such as beliefs and intentions through introspection. Such approaches already presuppose the private mind that the child must learn about rather than explain its development. In contrast, we argue that the mind must develop through social relations (Carpendale & Lewis, 2015).

Once children have developed a vocabulary for talking about human activity in psychological terms, then they can use this language to reflect on such activity, both their own and others. This makes an additional form of social understanding possible as well as the nonlinguistic, lived sensorimotor, practical skills, described above, that develops earlier in infancy and is assessed with so-called infant false-belief tasks. Such practical skills involve infants' developing expectations concerning what they can do with objects as well as their emerging understanding within social interaction such as how to make requests through learning to anticipate how others respond to their actions.

What does this approach imply for thinking about the neural bases for social understanding? The goal of studying the neural bases of mentalizing or social understanding is somewhat ambiguous. This can refer to discovering the bottom of or foundation for social understanding, on which the entirety of social cognition and its development depends. Clearly, neural activity is required for social understanding, but if the goal is to learn about social understanding by reducing it to and explaining it only in terms of neural activity, this is problematic because social understanding does not exist at that level of neural activity. The functioning of neurons depends on cellular chemistry that is required yet has little to do with social understanding. Such understanding becomes possible at the level of the experiencing being. Thinking does not happen just in the brain (Bennett & Hacker, 2003; Coulter, 2008; Malcolm, 1986). Instead, the nervous system mediates between the person and the world. "The nervous system transforms the physical energies so that from the wild dance of the photons there emerges the orderliness of the visible world" (Straus, 1963, p. 182). Neural biological systems are involved in setting up the interactional context in with infants can develop human forms of thinking, and these systems make such development possible (e.g., Decety, Bartal, Uzefovsky, & Knafo-Noam, 2016; Johnson, Jones, & Gliga, 2015). The activity of the nervous system enables a person to engage with the world, and neural pathways are shaped through experience in a form of "biographical biology" (Fuchs, 2011). Infants learn to anticipate the outcomes of their actions, which extends to learning to anticipate others' actions, establishing a lived form of early social understanding. This also extends to communication through learning how others respond to one's actions, resulting in the potential for intentional communication and for more complex forms of social understanding involving language. From this perspective, the neural bases for social understanding are not different from the functioning of the nervous system that makes it possible to mediate between the person and the world.

We are not criticizing the goal of identifying "neural correlates" of thinking about social matters. What we criticize is the interpretation of this neural activity. There is a strong tendency to think of that neural activity as the thinking itself, and further as computation or information processing. However, as Wittgenstein pointed out, "One of the most dangerous of ideas for a philosopher is, oddly enough, that we think with our heads or in our heads" (1967, § 605). Thinking does not just happen in the brain. It is a social process (e.g., Bennett & Hacker, 2003).

There will be parts of persons' brains that will be more active during such thinking. However, this does not mean that thinking happens just in the brain. Instead, the human nervous system enables people to engage with others in shared social routines on which communication is based and then thinking can be based. The fact that a dancer can imagine dance movements does not mean that the dancing is going on just in her head, or that the "neural correlates" are the dancing. Even though it takes two to tango, the dance can be imagined by one. In the process of learning to dance the tango, the human nervous system, and the body are all needed in addition to a dance partner in order to learn how to coordinate movements. This process is thoroughly biological at multiple levels from sensorimotor to neural as well as social with no way to clearly separate these somewhat artificial categories. Similarly, it takes two to converse, but once a child has learned a language, she can use this social process as a tool to think "in her head."

Understanding the point we make in this chapter requires making a shift in perspective, a Copernican shift, from conceptualizing thinking as in the brain and as the center of everything, to viewing the development of thinking as the outcome of an ongoing process of multilayered interactions within the human developmental system. The nervous system constitutes one layer within the developing system. The complexity of the human nervous system makes it possible to create a more distanced, mediated relation with the world ("mediated immediatedness," Plessner, 1928), which is the basis for the human form of life ("natural artificiality," Plessner, 1928). We have discussed the development of human forms of communication, and the resulting language makes thinking possible. Of course, this thinking does involve neural activity, but it is misleading to focus just on that aspect of the whole process. It is a developmental outcome of a social skill.

Trying to specify "the biological endowment" implies a separation of biology and environment. Instead, we could talk about the requirements that are necessary for the development of social understanding. Infants must be able to anticipate what will happen when they act in certain ways with respect to objects and people. This allows for the development of knowledge. It requires a nervous system and a system of muscles that enable the infant to engage with the world, making sensorimotor activity possible. The infant as an agent gradually learns to control her hands and arms through repeating actions that produced outcomes of interest. Through this process of repeating action patterns, the infant develops knowledge in the constructivist sense of learning the interactive potential of the world through acting on the world—learning to anticipate what will happen when she does certain things. This process of interaction is embedded in a social, emotional, cultural, and historical system within which human infants learn to communicate and to think.

Conclusion

We have examined how philosophical preconceptions influence research in social cognitive neuroscience, and in particular we have examined the implications of these worldviews in the case of claims that forms of thinking can be innate and that thinking is computation or information processing. We have argued that infants do not simulate others' experience based on their own experience, nor are they born with theories or formulate theories (e.g., Carpendale & Lewis, 2006, 2015). We have argued that communication does not work through encoding preexisting meaning into words that are transmitted to others and decoded, as in the code model. Instead, meaning is based on social relations. The view of social understanding as linked to communication and other aspects of social life fits with the findings that many brain regions tend to be active in thinking about social matters (Mar, 2011).

We have reviewed recent discussions in biology and developmental systems theory, suggesting that claims of innate knowledge and forms of thinking do not seem to be consistent with current views in biology. We have argued that the neural interconnectivity involved in interacting with others cannot be prespecified by genes and instead develops as part of the human developmental system. The human nervous system links the infant to the world and further interconnectivity develops after the child is born. This enables children to anticipate outcomes through experience. From a developmental systems approach, genetic influences are not discounted but instead are set within a system of multiple bi-directionally interacting factors, through which the effect of genes is modulated (e.g., Carpendale, Sokol, & Müller, 2010). There is no clear dichotomy between nature and nurture, or biological and social factors because they are too interwoven to separate, and they mutually create each other.

The task in this book is to explain the development of social understanding with a focus on the neural bases for this skill. For this, we need to account for how the world becomes meaningful for children and how they come to understand and think about the physical and social worlds within which they live through learning to talk about human activity in psychological terms. But meaning cannot be found by only studying the level of neurons firing (Straus, 1963). Supported by neural interconnectivity, human thinking is about the world and thus is meaningful. To explain this, we need to move to the level of the person coupled with the world and then consider the neural interconnectivity that is required for such engagement with the world. The human nervous system both enables interaction with the physical and social world and is also shaped through that interaction. As an alternative to the approaches we have criticized, we have sketched in a process-relational approach to the development of social understanding (Carpendale, Hammond, & Atwood, 2013; Carpendale & Lewis, 2015).

References

Baldwin, J. M. (1906). *Thoughts and things: Vol. 1. Functional logic*. New York, NY: The MacMillan Company.

Bates, E. (1999). Plasticity, localization and language development. In S. H. Broman & J. M. Fletcher (Eds.), *The changing nervous system: Neurobehavioral consequences of early brain disorders* (pp. 214–253). New York, NY: Oxford University Press.

Bates, E. (2005). Plasticity, localization, and language development. In S. T. Taylor, J. Langer, & C. Milbrath (Eds.), *Biology and knowledge revisited: From neurogenesis to psychogenesis* (pp. 205–253). Mahwah, NJ: Erlbaum.

Bates, E., Camaioni, L., & Volterra, V. (1975). The acquisition of performatives prior to speech. *Merrill-Palmer Quarterly, 21*(3), 205–226.

Bennett, M. R., & Hacker, P. M. S. (2003). *The philosophical foundations of neuroscience*. Oxford: Blackwell.

Bernstein, R. J. (2010). *The pragmatic turn*. Malden, MA: Polity.

Canfield, J. V. (2007). *Becoming human: The development of language, self, and self-consciousness*. New York, NY: Palgrave Macmillan.

Carpendale, J. I. M., Frayn, M., & Kurcharczyk, P. (2017). The social formation of human minds. In J. Kiverstein (Ed.), *Routledge handbook of the philosophy of the social mind* (pp. 189–207). New York, NY: Routledge.

Carpendale, J. I. M., Hammond, S. I., & Atwood, S. (2013). A relational developmental systems approach to moral development. In R. M. Lerner & J. B. Benson (Eds.), *Advances in Child Development and Behavior: Vol. 45. Embodiment and epigenesis: Theoretical and methodological issues in understanding the role of biology within the relational developmental system* (pp. 105–133). New York, NY: Academic Press.

Carpendale, J. I. M., & Lewis, C. (2006). *How children develop social understanding*. Oxford: Blackwell.

Carpendale, J. I. M. & Carpendale, A. B. (2010). The development of pointing: From personal directedness to interpersonal direction. *Human Development, 53*, 110–126.

Carpendale, J. I. M, & Lewis, C., (2015). The development of social understanding. In L. Liben & U. Müller (Vol. Eds.), R. Lerner (editor-in-chief), *Handbook of Child Psychology and Developmental Science: Vol. 2. Cognitive processes* (7th ed., pp. 381–424). New York, NY: Wiley Blackwell.

Carpendale, J. I. M., Sokol, B., & Müller, U. (2010). Is a neuroscience of morality possible? In P. Zelazo, M. Chandler, & E. Crone (Eds.), *Developmental social cognitive neuroscience* (pp. 289–311). New York, NY: Psychology Press.

Carpendale, J. I. M., & Ten Eycke, K. (2020). From reflex to meaning via shared routines. In M. F. Mascolo & T. Bidell (Eds.), *Handbook of integrative psychological development: Essays in honor of Kurt W. Fischer*. New York, NY: Routledge.

Chapman, M. (1999). Constructivism and the problem of reality. *Journal of Applied Developmental Psychology, 20*, 31–43.

Clark, R. A. (1978). The transition from action to gesture. In A. Lock (Ed.), *Action, gesture and symbol: The emergence of language* (pp. 231–257). New York, NY: Academic Press.

Coulter, J. (2008). Twenty-five theses against cognitivism. *Theory, Culture and Society, 25*(2), 19–32. https://doi.org/10.1177/0263276497086789

Decety, J., Bartal, I. B.-A., Uzefovsky, F., & Knafo-Noam, A. (2016). Empathy as a driver of prosocial behaviour: Highly conserved neurobehavioural mechanisms across species. *Philosophical Transactions of the Royal Society B, 371*, 1–11.

Dewey, J. (1960). *On experience, nature, and freedom*. New York, NY: The Bobbs-Merrill Company, Inc..

Di Paolo, E., & De Jaegher, H. (2012). The interactive brain hypothesis. *Frontiers in Human Neuroscience, 6*(163), 1–16.

Farroni, T., Massaccesi, S., Pividori, D., & Johnson, M. H. (2004). Gaze following in newborns. *Infancy, 5*, 39–60.

Fisher, S. E. (2006). Tangled webs: Tracing the connections between genes and cognition. *Cognition, 101*, 270–297.

Fuchs, T. (2011). The brain—A mediating organ. *Journal of Consciousness Studies, 18*, 196–221.

German, T. P., & Leslie, A. M. (2004). No (social) construction without (meta) representation: Modular mechanisms as a basis for the capacity to acquire an understanding of mind. *Behavioral and Brain Sciences, 27*, 106–107.

Goldberg, B. (1991). Mechanism and meaning. In J. Hyman (Ed.), *Investigating psychology: Sciences of the mind after Wittgenstein* (pp. 48–66). New York, NY: Routledge.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin, 138*, 1085.

Gordon, R. M. (1986). Folk psychology as simulation. *Mind & Language, 1*, 156–171.

Gottlieb, G. (2007). Probabilistic epigenesis. *Developmental Science, 10*, 1–11.

Hanson, N. R. (1958). *Patterns of discovery: An inquiry into the conceptual foundations of science*. Cambridge: Cambridge University Press.

Harris, P. L. (1991). The work of the imagination. In A. Whiten (Ed.), *Natural theories of mind* (pp. 283–304). Oxford: Blackwell.

Heil, J. (1981). Does cognitive psychology rest on a mistake? *Mind, 90*, 321–342.

Heil, J. (1998). *Philosophy of mind*. New York, NY: Routledge.

Hobson, P. (2002). *The cradle of thought: Explorations of the origins of thinking*. London: Macmillan.

Jablonka, E., & Lamb, M. J. (2014). *Evolution in four dimensions: Genetic, epigenetic, behavioral, and symbolic variation in the history of life* (Rev. ed.). Cambridge, MA: MIT Press.

Johnson, M. H., Jones, E. J. H., & Gliga, T. (2015). Brain adaptation and alternative developmental trajectories. *Development and Psychopathology, 27*, 425–442.

Jopling, D. (1993). Cognitive science, other minds, and the philosophy of dialogue. In U. Neisser (Ed.), *The perceived self* (pp. 290–309). Cambridge, MA: MIT Press.

Kenny, A. (1991). The homunculus fallacy. In J. Hyman (Ed.), *Investigating psychology: Sciences of the mind after* (pp. 155–165). London: Routledge. (original work published 1971)

Kiverstein, J., & Miller, M. (2015). The embodied brain: Towards a radical embodied cognitive neuroscience. *Frontiers in Human Neuroscience, 9*(237), 1–11.

Kobayashi, H., & Kohshima, S. (2001). Unique morphology of the human eye and its adaptive meaning: Comparative studies of external morphology of the primate eye. *Journal of Human Evolution, 40*, 419–435.

Lehrman, D. S. (1953). Problems raised by instinct theories. *The Quarterly Review of Biology, 28*, 337–365.

Lerner, R. M., Agans, J. P., DeSouza, L. M., & Hershberg, R. M. (2014). Developmental science in 2025: A predictive review. *Research in Human Development, 11*, 255–272.

Lewontin, R. C. (2001). Gene, organism and environment. In S. Oyama, P. E. Griffiths, & R. D. Gray (Eds.), *Cycles of contingency: Developmental systems and evolution* (pp. 55–66). Cambridge, MA: The MIT Press. (Original work published 1983).

Lickliter, R., & Honeycutt, H. (2015). Biology, development, and human systems. In W. F. Overton & P. C. M. Molenaar (Vol. Eds.), R. Lerner (editor-in-chief), *Handbook of Child Psychology and Developmental Science: Vol. 1. Theory and method* (7th ed.). New York, NY: Wiley Blackwell.

Lickliter, R., & Witherington, D. C. (2017). Towards a truly developmental epigenetics. *Human Development, 60*, 124–138.

Malcolm, N. (1986). *Nothing is hidden: Wittgenstein's criticism of his early thought*. Oxford: Basil Blackwell.

Mameli, M., & Bateson, P. (2006). Innateness and the sciences. *Biology and Philosophy, 21*, 155–188.

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology, 62*, 103–134.

Mareschal, D., Johnson, M. H., Sirois, S., Spratling, M. W., Thomas, M. S. C., & Westermann, G. (2007). *Neuroconstructivism: How the brain constructs cognition* (Vol. 1). New York, NY: Oxford University Press.

McDonough, R. (1989). Towards a non-mechanistic theory of meaning. *Mind, 98*, 1–21.

McDonough, R. (1999). Bringing cognitive science back to life. *Idealistic Studies, 29*, 173–214.

Mead, G. H. (1934). *Mind, self and society: From the standpoint of a social behaviorist*. Chicago, IL: University of Chicago Press.

Meaney, M. J. (2010). Epigenetics and the biological definition of gene x environment interactions. *Child Development, 81*, 41–79.

Meltzoff, A. N. (2007). 'Like me': A foundation for social cognition. *Developmental Science, 10*, 126–134.

Meltzoff, A. N., Gopnik, A., & Repacholi, B. M. (1999). Toddlers' understanding of intentions, desires, and emotions: Explorations of the dark ages. In P. D. Zelazo, J. W. Astington, & D. R. Olson (Eds.), *Developing theories of intention* (pp. 17–41). Mahwah, NJ: Lawrence Erlbaum Associates.

Müller, U., & Carpendale, J. I. M. (2004). The development of social understanding in infancy. In J. I. M. Carpendale & U. Müller (Eds.), *Social interaction and the development of knowledge* (pp. 215–238). Mahwah, NJ: Erlbaum.

Onishi, K., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*, 255–258.

Overton, W. F. (2015). Processes, relations, and relational-developmental-systems. In W. F. Overton & P. C. M. Molenaar (Vol. Eds.), R. Lerner (editor-in-chief), *Handbook of Child Psychology and Developmental Science: Vol. 1. Theory and method* (7th ed., pp. 9-62). New York, NY: Wiley Blackwell.

Oyama, S., Griffiths, P. E., & Gray, R. D. (2001). Introduction: What is developmental systems theory? In S. Oyama, P. E. Griffiths, & R. D. Gray (Eds.), *Cycles of contingency: Developmental systems and evolution* (pp. 1–11). Cambridge, MA: The MIT Press.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: The MIT Press.

Piaget, J. (1952). *The origins of intelligence in children*. New York, NY: International Universities Press. (Original work published in 1936).

Piaget, J. (1970). Piaget's theory. In P. Mussen (Ed.), *Carmichael's manual of child psychology* (3rd ed., pp. 703–732). New York, NY: Plenum Press.

Pinker, S. (1997). *How the mind works*. New York, NY: W. W. Norton & Company.

Plessner, H. (1928). *Die Stufen des Organischen und der Mensch* [The levels of the organic and the human being]. Berlin: Walter de Gruyter.

Portmann, A. (1990). *A zoologist looks at humankind*. New York, NY: Columbia University Press. (Original work published 1944).

Putnam, H. (1988). *Representation and reality*. Cambridge, MA: The MIT Press.

Reddy, V., Markova, G., & Wallot, S. (2013). Anticipatory adjustments to being picked up in infancy. *PLoS One, 8*, 1–9.

Rodríguez, C., Moreno-Núñez, A., Basilio, M., & Sosa, N. (2015). Ostensive gestures come first: Their role in the beginning of shared reference. *Cognitive Development, 36*, 142–149. https://doi.org/10.1016/j.cogdev.2015.09.005

Sameen, N., Thompson, J., & Carpendale, J. I. M. (2013). Further steps toward a second-person neuroscience. *Behavioral and Brain Sciences, 36*, 437.

Scheler, M. (1954). *The nature of sympathy* (P. Heath, Trans.). Hamden, CT: Archon Books. (Original work published 1913).

Senju, A., & Johnson, M. H. (2009). The eye contact effect: Mechanisms and development. *Trends in Cognitive Sciences, 13*, 127–134.

Service, V., Lock, A., & Chandler, P. (1989). Individual differences in early communicative development: A social constructivist perspective. In S. von Tetzchner, L. S. Siegal, & L. Smith (Eds.), *The social and cognitive aspects of normal and atypical language development* (pp. 21–49). New York, NY: Springer.

Shanker, S. (1998). *Wittgenstein's remarks on the foundations of AI*. New York, NY: Routledge.

Sherry, A. E., & Saxe, R. (2016). What neuroscience can reveal about cognition and its origins. In D. Barner & A. S. Baron (Eds.), *Core knowledge and conceptual change*. New York, NY: Oxford University Press.

Spencer, J. P., Blumberg, M. S., McMurray, B., Robinson, S. R., Samuelson, L. K., & Tomblin, J. B. (2009). Short arms and talking eggs: Why we should no longer abide the nativist–empiricist debate. *Child Development Perspectives, 3*, 79–87.

Stiles, J. (2009). On genes, brains, and behavior: Why should developmental psychologists care about brain development? *Child Development Perspectives, 3*, 196–202.

Stiles, J., Brown, T. T., Haist, F., & Jernigan, T. L. (2015). Brain and cognitive development. In L. Liben & U. Müller (Vol. Eds.), R. Lerner (editor-in-chief), *Handbook of Child Psychology and Developmental Science: Vol. 2. Cognitive processes* (7th ed., pp. 9–62). New York, NY: Wiley Blackwell

Straus, E. (1963). *The primary world of the senses: A vindication of sensory experience*. London: The Free Press of Glencoe.

Tomasello, M., Hare, B., Lehmann, H., & Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: The cooperative eye hypothesis. *Journal of Human Evolution, 52*, 314–320.

Tooby, J., & Cosmides, L. (1995). Forward. In S. Baron-Cohen, *Mindblindness: An essay on autism and theory of mind* (pp. xi–xviii). Cambridge, MA: MIT Press.

Winch, P. (1958). *The idea of a social science and its relation to philosophy*. London: Routledge and Kegan Paul.

Wittgenstein, L. (1967). Zettel (edited by G. E. M. Anscombe and G. H. Wright, translated by G. E. M. Anscombe). Oxford: Basil Blackwell.

Zahavi, D. (2008). Simulation, projection and empathy. *Consciousness and Cognition, 17*, 514–522.

Part IV
Cognitive Components

The Tree of Social Cognition: Hierarchically Organized Capacities of Mentalizing

Bertram F. Malle

Mental state inference, theory of mind, mentalizing—all these terms denote the capacity to represent something beyond, behind, or simply different from physical objects, moving bodies, and expressive faces. Scholars of philosophy have for thousands of years pondered how "mind" works; psychology brought scientific methods to such investigations. A few scholars in the twentieth century then discovered that not only they themselves but ordinary humans, too, wonder about the mind; and it became clear that such mind wondering underlies and enables social interaction, culture, and morality, as much as politics, religion, and technology.

The emerging picture is that, in response to intense demands of social group living, human beings have evolved a number of capacities that allow them to make sense of other agents—to interpret, explain, and predict their behavior, share their experiences, and coordinate interactions with them (e.g., Bloom, 2007; Dunbar & Shultz, 2007; Tomasello, 1998). These enabling capacities include simpler processes such as face detection or mimicry; complex processes such as imaginative simulation and mental state inference; and fundamental concepts such as *intentionality* and *belief*. The diversity of these capacities (Malle, 2008; Mitchell, 2006) and their different ways and degrees of representing mental states (Apperly & Butterfill, 2009; Poulin-Dubois, Brooker, & Chow, 2009; Sterck & Begeer, 2010) require a more inclusive term than *theory of mind* or *mentalizing*. I suggest that these capacities are best subsumed under the broader label *social cognition.* These social-cognitive capacities belong together, not because they form a "module" or can somehow be localized in a particular brain area; rather, what unites them is their responsiveness to other intentional agents and the benefits they convey when interacting with those agents. My investigation of mentalizing is thus contextualized

B. F. Malle (✉)
Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, USA
e-mail: bfmalle@brown.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_17

within a broad framework of social cognition, which I introduce as a hierarchically organized structure.

A Broad Framework: The Tree of Social Cognition

How are the capacities of social cognition related to each other? I propose that the structure of social cognition is hierarchical, ranging from lower-order to higher-order capacities. This hierarchy appears in at least three ways: (a) lower-order capacities (LC) develop earlier in life and are likely to have evolved earlier in human history than higher-order capacities (HC); (b) LC have lower processing demands than HC and may only weakly rely on explicit mind representations; and (c) LC are often inputs to or even requirements for HC. I am not proposing an LC-HC dichotomy but rather a multi-layered hierarchical structure: a tree of social cognition (Malle, 2015).

Figure 1 displays the approximate hierarchy of capacities of social cognition, starting at the bottom with the fundamental identification of agents in the environment and building from simpler processes of gaze following to the complex processes of mental state and trait inferences. The tree is not a comprehensive representation of all social-cognitive capacities, and the exact location of any given tool is imprecise and debatable. However, the evidence for an overall hierarchy is rather compelling, exemplified by evidence on orderings in development (Poulin-Dubois et al., 2009; Sirois & Jackson, 2007; Wellman, Cross, & Watson, 2001),

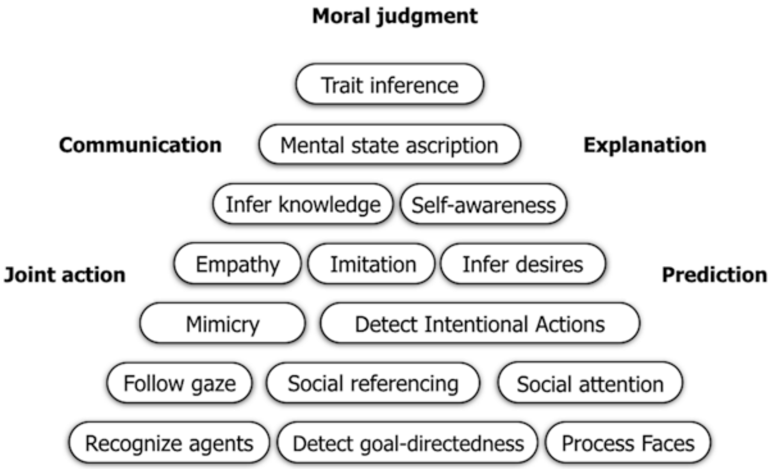


Fig. 1 The tree of social cognition. In the bottom layers, we find lower-order capacities (earlier-developing and more likely to be present in nonhuman animals) that facilitate higher-order capacities in the upper layers. Bundles of capacities also enable more complex social-cognitive activities such as explanation, communication, and moral judgment. (This is a revised version of Figure 12.2 in Voiklis & Malle, 2017)

evolution (Call & Tomasello, 2008; Povinelli & Preuss, 1995), and cognitive processing (Malle & Holbrook, 2012; Van Overwalle, Van Duynslaeger, Coomans, & Timmermans, 2012).[1]

Figure 1 also displays, outside the tree, important activities that are enabled by combinations of social-cognitive (and other) capacities. For example, *communicating* with others involves at least the basic tools of gaze following and joint attention to understand linguistic reference, as well as speakers' inferences of what the listener already knows, does not want to hear, or tries to find out (Barker & Givón, 2005; Clark, 1996; Krauss & Fussell, 1991). Likewise, there is no doubt that *explaining* human behavior relies on the careful scrutiny of gaze and attention to infer intentionality, and on inferences of specific desires, knowledge, and more complex mental states (Malle, 2004). And *moral judgment* takes into account not just observed behavior and outcomes but the subtleties of intentionality, the agent's reasons, and what the agent should and could have known (Alicke, Buckingham, Zell, & Davis, 2008; Cushman, 2008; Malle, Guglielmo, & Monroe, 2014). These activities are important for a broader picture of social cognition and social interaction, but they are not the focus of this chapter.

In what follows I will discuss each of the depicted capacities and offer evidence in favor of their approximate location within the tree. This evidence will come to a significant extent from developmental research, which offers the richest currently available database, and also from some comparative work and adult cognitive and social psychology. My main goal is to show the diversity of ways in which "mentalizing" can occur—ways in which humans connect to other minds.

Capacities of Social Cognition

Agents

A foundational task in social life is to recognize objects in the world that are candidates for having minds: *agents*. A few features can turn an object into a candidate agent: having eyes, acting contingently (responsive turn taking), and self-propelled behavior with equifinality (i.e., continuing pursuit of the goal under changing conditions) (Johnson, Slaughter, & Carey, 1998; Luo & Choi, 2013; Premack, 1990). Part of what elicits perceived agency is biological motion, which already 3–5 month-old infants can identify (Moore, 2011) and which grows into a sophisticated bottom-up/top-down integrative body perception (Johnson & Shiffrar, 2013). However, even

[1] Despite the boundaries I drew around the capacities in Fig. 1, I do not assume that each of them has its own "circuit." A capacity here is really the pattern of performance of certain functions under certain conditions, and currently we do not know how distinct the computations and neural substrates are for these functions. In fact, because I argue that many capacities build on each other, I expect a smaller number of divisible substrates that differentiate and recombine to enable interconnected performances.

nonbiologically moving objects (fury blobs, boxes, or triangles) are treated as agents, by children and adults alike, when they exhibit the features of equifinality (Heider & Simmel, 1944; Johnson, Shimizu, & Ok, 2007; Király, Jovanovic, Prinz, Aschersleben, & Gergely, 2003), as long as the observed movement is continuous (Berry, Kean, Misovich, & Baron, 1991). Treating certain entities as agents is a prerequisite for making further inferences about those entities' minds or moral status, even in the case of robots (e.g., Fiala, Arico, & Nichols, 2014; Gray & Wegner, 2012). Some of those inferences are so intimately connected to agency that they both provide evidence for agent status and are expected to be present once agent status is granted, including goal-directedness and gaze following, discussed next.

Goal-Directedness

A particularly robust recognition of agents relies on detecting a behavior's goal-directedness. Within the first year of life, infants show a sensitivity to agents' coordinated movements toward objects (Wellman & Phillips, 2001; Woodward, 1998). Equifinality appears to be the most diagnostic cue in those movements (Luo & Choi, 2013) and has been recognized as a fundamental element of the adult conception of intentionality (Heider, 1958). Appreciating goal-directedness is not itself a mental state inference but a sophisticated theory of behavior; it guides the perceiver's attention to certain patterns of behaviors (e.g., reaching, looking) by certain kinds of entities ("agents") and builds expectations about future behaviors by these entities. For example, when 6- to 9-month-old infants see a human arm repeatedly reach for an object, they expect it to continue to reach for that object even when the object changes location; but they do not expect this pattern of object-directedness from a mechanical claw (Woodward, 1998). A few months later, infants understand that even just gaze behavior (without a reach) can indicate the same object-directedness (Woodward, 2003). Reach and gaze are of course diagnostic of desires, interest, and other mental states; infants thus carefully track the kinds of behavior patterns that guide them toward the minds of others even before they fully understand those minds.

Faces

Faces capture and maintain 6-month-old infants' attention, but not 3-month-olds' (Di Giorgio, Turati, Altoè, & Simion, 2012). After 7 months, infants are sensitive to point-light displays of dynamic facial expressions such as surprise (Ichikawa, Kanazawa, Yamaguchi, & Kakigi, 2010), and infants' brains differentiate between familiar and unfamiliar faces (de Haan & Nelson, 1999) and between happy and fearful faces (Jessen & Grossmann, 2015). In adults, the brain differentiates familiar from unfamiliar faces between 140 and 200 ms after exposure (Barragan-Jason, Cauchoix, & Barbeau, 2015), and a conscious recognition response is possible after just over 300 ms (Ramon, Caharel, & Rossion, 2011). However, more differentiated judgments, such as recognizing specific emotions, takes considerably longer (e.g., Dodonova & Dodonov, 2012).

Artists are aware of the power of face and eyes, and of contingent and equifinal behavior, as those features constitute the vocabulary to make inanimate objects come alive (Lundmark, 2017; Thomas & Johnston, 1995). They also provide the foundation for further social-cognitive skills, discussed next, that can develop only because of the infant's keen attention to those foundational features.

Gaze Following

Following other agents' direction of attention is a powerful tool to learn about the world, about its treasures and its threats. A basic, perhaps reflexive ability to follow a body, head, and gaze has been found in several mammals, even birds (Kehmeier, Schloegl, Scheiber, & Weiß, 2011), and in infants from at least 6 months of age (Gredebäck, Astor, & Fawcett, 2018). More sophisticated gaze following involves a rudimentary idea of *seeing* as a mind-world connection: by about 11 months, infants selectively follow open eyes but not closed eyes (Brooks & Meltzoff, 2005), a distinction that may be too difficult for chimpanzees (Povinelli & Eddy, 2000). Inferring which object a person is looking at when the possible objects of attention are more numerous, partially occluded, or spatially more diverse comes online a little later, at about 14 months (Carpenter, Nagell, & Tomasello, 1998; Slaughter & McConnell, 2003). And only with additional maturation do children seem to interpret looking as an internal state that can express intention even in the absence of an object of interest (Moore & Povinelli, 2007).

We see here, as in other capacities, that gaze following undergoes development and refinement, from a more behavioral to a more mentalistic processing level. This mix of behavioral and mentalistic processing is apparent in adult behavior as well, enriched by social impact. A single person on the street looking up at a sixth-floor window induces over 40% of people to look up as well; two people looking up persuade 60%; and five entice 80% (Milgram, Bickman, & Berkowitz, 1969). This response is induced by a behavioral trigger of another's looking behavior, but it goes beyond an orienting reflex; it includes a consideration of the diagnosticity of the gaze behavior: If an increasing number of people look up, they must have a reason. We wonder about what is up there but also why so people are interested in it.

Social Referencing

Not only is another's gaze a useful piece of information, but the person's facial expression can indicate whether the attended object should be valued or not. Social referencing is the act of using such diagnostic information about an object's valence, significance, or meaning (Klinnert, Campos, Sorce, Emde, & Svejda, 1983). By 10–12 months, children begin to decode such facial reactions about the value of objects (Slaughter & McConnell, 2003; Walden & Ogan, 1988); and a little later they care about the specific object that the adult attends to, not another one nearby (Moses, Baldwin, Rosicky, & Tidball, 2001). There is also evidence that infants not only passively use social appraisals but actively seek them. Such seeking behavior was shown in a classic developmental study, where 12-month-olds looked to their caregiver to help interpret a potential threat (visual cliff) and crossed only when the caregiver emoted a positive attitude (Sorce, Emde, Campos, & Klinnert, 1985). The active search for information in others' behaviors, emotions, and attitudes when in ambiguous situations continues to be important in adulthood (Walle, Reschke, & Knothe, 2017), such as in the classic studies on bystander intervention, where the search for information is apparent but may lie below people's own awareness (Latané & Darley, 1968). More broadly, social referencing can be seen as the foundation for conformity, but it goes beyond mere copying of *behavior* to the adoption of the social partners' *interpretation* of the situation (Feinman, 1992).

Social Attention

The maturing of social attention management from simpler, less mentalistic to more complex, mentalistic variants is visible in the increasingly sophisticated pointing behavior of 9–18-month-olds (Franco, 2005). In "imperative pointing," the child uses pointing gestures to express a desire for an object to another person (demonstrated before 12 months). Declarative pointing is intended to shift the other's attention toward an object (when the other is not yet aware of the object), and a majority of children do it by 12–15 months. Yet more sophisticated is coordinated joint attention, involving alternating gaze between the object of interest and the other person. This capacity emerges by about 15 months (Bakeman & Adamson, 1984), although first age of onset may be earlier (Carpenter, Nagell, & Tomasello, 1998). Human-reared chimpanzees do not seem to show such active joint attention (Tomasello & Carpenter, 2005). Active attentional engagement is a powerful prerequisite for learning, broader collective intentionality, and culture (Tomasello & Rakoczy, 2003). More generally, the sharing of experience strengthens memory (Hoerl & McCormack, 2005) and is psychologically rewarding (Higgins & Pittman, 2008).

Intentionality

We have seen that detecting goal-directedness is a basic and early developing capacity; how is detecting intentionality different? To continue our theme, the former requires little to no consideration of a mind (but rather relies on recognizing certain systematic behavior patterns); the intentionality detection does require such consideration. For one thing, any social perceiver faces the challenge that behavior usually comes in continuous streams, so the most important intentional actions must be extracted from the stream. Already at 12 months, infants show sensitivity to the surface features that characterize intentional actions (e.g., timing, contact, direction of attention) and are able to recognize the points at which intended actions are completed (Baldwin & Baird, 2001; Saylor, Baldwin, Baird, & LaBounty, 2007). Between 14 and 18 months, they can make the categorical distinction between intentional and unintentional behaviors (Carpenter, Akhtar, & Tomasello, 1998), which is also available to chimpanzees (Call & Tomasello, 1998). In human adults, these basic intentionality judgments differentiate conceptually and are made along two paths, depending on available information and judgment demands. Along the "slow and measured" path, people take into account what they know about the agent's context, mental states, and so on. For example, when we wonder whether a colleague who made a hurtful remark did it intentionally, we consider whether he holds a grudge, knew about our vulnerability, was aware of what he was actually saying, etc. (Malle & Knobe, 1997). Along the "fast and configural" path, many observed behaviors simply "look" intentional, and these configurations are well learned from numerous experiences of one's own and others' actions. These configurations allow intentionality judgments to be made faster than other mental state judgments (Decety & Cacioppo, 2012; Malle & Holbrook, 2012), and some of them are encoded as prototypes into action verbs with a strong intentionality implication (e.g., *reach*, *walk*, *look*, *help*; Malle, 2002).

Mimicry

Humans show some degree of synchronization at fairly low levels, such as heart rate, muscle tension, and pupil dilation (Prochazkova & Kret, 2017), but most useful for social cognition is mimicry of movements, postures, and gestures, because they can confer and reflect socially affiliative behavior (Chartrand & Bargh, 1999; Leighton, Bird, Orsini, & Heyes, 2010). Also important is mimicry of emotional expressions in the face, because it can facilitate shared emotions. A well-known proposal of mimicry in newborns (Meltzoff & Moore, 1977, 1997) has been challenged both at the level of evidence and interpretation (Jones, 2009; Keven & Akins, 2017; Oostenbroek et al., 2016; Ray & Heyes, 2011; Vincini, Jhang, Buder, & Gallagher, 2017). A systematic study of a range of behaviors across a range of ages (6–20 months) showed no above-chance mimicry at 6 months but increasing mimicry between 12 and 18 months, varying by specific behavior (Jones, 2007). In natural play interactions, 16 month-old toddlers begin to mimic each other and increase such behavior steadily over their development (Eckerman, Davis, & Didow, 1989; Nadel, 2002). One study suggests that chimpanzees and gorillas mimic each other's facial expressions (Palagi, Norscia, Pressi, & Cordoni, 2019), but the specific situation (play fighting) may represent a third variable that causes similar expressions in both animals. However, some evidence exists for simple mimicry among nonhuman primates, such as contagious yawning (Campbell & de Waal, 2014) or entrainment of finger tapping and other simple motor behaviors (Yu, Hattori, Yamamoto, & Tomonaga, 2018).

Though mimicry is often taken to be an automatic, inevitable mechanism (Heyes, 2011), a good deal of mimicry in humans appears to be regulated, or at least modulated, by higher-order processes. Looking at facial mimicry, which appears to be both spontaneous and fast (Sato & Yoshikawa, 2007), we find that the copying behavior is far too socially strategic to be left to "mirror neurons" (Fischer & Hess, 2017; Wang & Hamilton, 2012). That is, emotion mimicry is more likely to appear when there is already a social connection to the other—e.g., through liking (Blocker & McIntosh, 2016) or shared group membership (Rauchbauer, Majdandžić, Stieger, & Lamm, 2016)—or when such a connection is desired—e.g., when one seeks to repair group integration (Cheung, Slotter, & Gardner, 2015). Thus, the lower-level capacity is integrated into the higher-level project of social regulation (Hess & Fischer, 2013).

Inferring Desire

Inferring desires is more demanding than recognizing the goal-directedness of behaviors (such as reaching for an object; Woodward, 1998), and it also goes beyond the category distinction of intentional vs. unintentional behavior. It involves representing the content of a desire "in" a person's mind. One way in which this representation manifests is through recognizing that another person can have a desire different from one's own, an ability that may emerge at 18 months (Repacholi & Gopnik, 1997), though subsequent replication attempts found evidence no earlier than 24 months (Ruffman, Aitken, Wilson, Puri, & Taumoepeau, 2018). As mental state verbs of desire appear in children's speech around 18 months (Bartsch & Wellman, 1995), we can safely say that children begin to master desire inferences sometime in their second year (Wellman & Woolley, 1990).

Advanced desire inferences are grounded not just in obvious behavioral cues (e.g., reaching movement) but in observed emotional reactions and, somewhat later, in observed eye gaze and pointing (Lee, Eskritt, Symons, & Muir, 1998). Desire inferences should also be apparent when the goal is not directly visible. Meltzoff (1995) suggested that 18-month-olds can infer what goal an actor is trying to achieve (e.g., in manipulating a novel object) and perform the kinds of actions that achieve the inferred goal. The exact mix of behavioral cues, object affordances, and inferred mental states is difficult to determine, but it is clear that there is a difference between analyzing behavior patterns that fairly directly reveal "corresponding" mental content (reaching for X = being directed to X) and analyzing behavior patterns that require some inference to reveal "noncorresponding" mental content (doing X → must want Y). Looking behind the observable and the obvious is of course the strength of sophisticated mental state inference—which later allows people to see through self-presentation, irony, and deception.

Imitation

Imitation is more involved than mimicry as it capitalizes on the newly gained ability to infer desires. In imitation, the perceiver attends to the other performing a novel behavior or manipulating an object in a novel way but with a particular desire or intention. The imitator then reproduces not just the observed behaviors but implements the inferred intentions or goals. Earliest evidence for such inference-based social imitation is found at 15–18 months of age (Johnson, Booth, & O'Hearn, 2001; Meltzoff, 1995). The evidence for this kind of imitation in nonhuman primates is suggestive but inconsistent and debated (Carrasco, Posada, & Colell, 2009; Herrmann, Call, Hernàndez-Lloreda, Hare, & Tomasello, 2007; Persson, Sauciuc, & Madsen, 2018; Subiaul, Renner, & Krajkowksi, 2016).

Imitation undergoes several developmental stages: Two-year-olds show the ability to infer a model's goal and copy only the relevant behaviors to achieve that goal (and not the ones that lead to failure). However, 3-year-olds start showing what is called overimitation (Lyons, Young, & Keil, 2007), as they also copy a model's *failed* attempts (Huang, Heyes, & Charman, 2006) and causally irrelevant behaviors (which other primates never do; Clay & Tennie, 2018). These patterns are robust across a number of cultural communities (Nielsen, Mushin, Tomaselli, & Whiten, 2014; Nielsen & Tomaselli, 2010), though they may appear later in some (Hewlett, Berl, & Roulette, 2016). Children have not lost their ability to distinguish intentional from unintentional behaviors, as they overimitate only intentional behaviors (Lyons, Damrosch, Lin, Macris, & Keil, 2011). Their detailed mimicking of new behaviors may represent an openness to learn novel skills, unusual social norms and rituals, and thus to affiliate with members of their community (Nielsen, 2018; Wen, Herrmann, & Legare, 2016). Indeed, overimitation is more likely in the presence of a social audience (Marsh, Ropar, & Hamilton, 2019), is more strongly triggered by the behavior of ingroup members (Gruber, Deschenaux, Frick, & Clément, 2019), and is sustained even in adult years (Flynn & Smith, 2012; Hewlett et al., 2016).

Empathy/Emotion Matching

There is a bundle of terms that refer to some form of emotional reaction to another person's emotions: empathy, empathic concern, sympathy, and emotional contagion. I will focus here on empathy, understood as having the same emotion as another person because one observes the other's emotion (Feshbach & Roe, 1968). This is similar to emotional contagion (Hatfield, Cacioppo, & Rapson, 1994), but additional cognitive and motivational mechanisms may facilitate or moderate the contagious response (e.g., Cameron et al., 2019).

Negative reactions to another's distress emerge early in infancy, but the mechanisms do not meet the adopted definition of empathy. Genuine empathy requires a differentiated matching of experienced emotions with observed emotions—so that the perceiver experiences E_1 when observing E_1, E_2 when observing E_2, etc. Contagious crying in newborns (if contagion at all; Ruffman, Lorimer, & Scarf, 2017) and 1–2 year-olds' concern for others' distress (Roth-Hanania, Davidov, & Zahn-Waxler, 2011) are relatively undifferentiated; no matter what the other's specific distress is, the perceiver has a general response of concern—which corresponds rather to sympathy, a particular emotion felt in response to a large variety of negative states in the other person. Likewise, nonhuman primates show consolation behavior, and if regarded as an emotional response (de Waal & Preston, 2017), it could be counted as sympathy.

The age of emergence of genuine empathic emotion matching is under debate, but even skeptics' results (Ruffman, Then, Cheng, & Imuta, 2019) suggest that in the second year of life, happy and sad videos lead to differential emotional responses on the happy-to-sad dimension. However, in that study and other studies on the same age group (e.g., Scambler, Hepburn, Rutherford, Wehner, & Rogers, 2007), happy stimuli elicited far stronger matching responses than sad stimuli. With age, this asymmetry declines somewhat. About half of 3- to 5-year-olds showed increasing sadness expressions to a video story when it moved to the sad climax (Stiles, 1985). And among 6–7 year-olds, both happy and sad story sequences led to high rates of matching self-reported emotions; however, corresponding rates were low for anger or fear (Feshbach & Roe, 1968). Likewise, among adults, happy and sad faces elicit happy and sad feelings, respectively, whereas anger, fear, and disgust at best do so inconsistently (Blairy, Herrera, & Hess, 1999; Hess & Blairy, 2001). These results suggest that perceivers do not simply "catch" emotions by mimicking the parallel emotional expression. The age-dependent conceptual interpretation of the emotion is necessary to perceive and replay the correct emotion; and some emotions are better matched than others, whether due to difficulty or motivation.

In adults, genuine and specific empathy occasionally results from mimicry, such as in studies that expose perceivers to extended dynamic video stimuli or a live interaction partner (Stel, Van Baaren, & Vonk, 2008; Stel & Vonk, 2010). However, even when mimicry emerges, it typically does not cause matching emotions (Blairy et al., 1999; Hess & Blairy, 2001). Empathy can come about through other means, such as hearing an emotional tone of voice (Neumann & Strack, 2000), imagining the other's emotion (Hawk, Fischer, & Van Kleef, 2011), or simulating the mere idea of an emotion (Hess, Houde, & Fischer, 2014). And empathy can be moderated by self-regulation (Hodges & Klein, 2001; Ochsner, 2013; Powell, 2018). It seems that actual emotional contagion is relatively rare, and empathy as emotion matching (through contagion or not) is a more learned, refined, and regulated response.

Inferring Knowledge

Between 18 and 30 months, children recognize that talking to someone or a nodding gesture can transmit knowledge from one person to another (Fusaro & Harris, 2013; Song, Onishi, Baillargeon, & Fisher, 2008). Children themselves also begin to use gestures to transmit information (Begus & Southgate, 2012) and pose a large number of questions (Chouinard, 2007) to seek information. By the middle of the third year, they selectively provide information, verbally or nonverbally, to others who don't know that information (O'Neill, 1996). Their language use, too, begins to reflect their emerging understanding of affirmed, denied, and requested knowledge (Harris, Yang, & Cui, 2017).

Handling such knowledge transfer is a critical capacity that not only conceptualizes mental states of knowing but separates knowledge as information from the minds that hold that knowledge. This allows perceivers to distinguish between people who know and those who don't know (Koenig & Harris, 2005) and to guide their social interactions by such differences. This in turn explains why children ask adults about food but other children about toys (Van der Borght & Jaswal, 2009), and it also enables us to do fine knowing very little about many things, as long as we know who in our community knows (Sloman & Fernbach, 2017).

Handling rapidly shifting belief and knowledge inferences is also critical in conversation, both to make subtle linguistic decisions (e.g., about "a" vs. "the"; Barker & Givón, 2005) and to tailor utterances to one's conversation partner, taking into account what they know, see, and hear (Fukumura, 2015; Krauss & Fussell, 1991).

Self-Awareness

Proto forms of self-awareness occur when infants' experiences of their own actions become models for understanding the actions of others (Sommerville, Woodward, & Needham, 2005), and several theorists would argue that experiences of one's own mental states are models for understanding the mental states of others (Goldman, 2009; Gordon, 1986). Evidence for the development of self-awareness is typically associated with body self-recognition in the famous mirror test (Gallup, 1970), in which the agent has to recognize themselves in the mirror by touching a mark on their own body (rather than on the mirror surface). Between 18 and 24 months, a majority of children pass the test, and many chimpanzees do too (Povinelli, Rulf, Landau, & Bierschwale, 1993).

However, being aware of one's present (bodily) state is one thing (Suddendorf & Butler, 2013); bridging one's past and present selves is more challenging. When children watched a video of themselves in which the experimenter put a sticker on their forehead, only a quarter of 2- and 3-year-olds immediately checked their forehead for the sticker, whereas three fourths of 4-year-olds did (Povinelli, Landau, & Perilloux, 1996). Such time-extended self-awareness emerges only slowly. Three-year-olds have trouble recognizing that their own past (false) beliefs actually motivated their own actions (Atance & O'Neill, 2004). Four- to 5-year-olds who were just taught some novel facts normally do not realize that they didn't know those facts a little earlier (Taylor, Esbensen, & Bennett, 1994). And only after 5 years of age can children report what they were thinking a short while ago (Flavell, Green, & Flavell, 1995; Louca-Papaleontiou, Melhuish, & Philaretou, 2012).

Awareness of the present moment is easier. Three- to 4-year-olds can recall a concrete false belief they had just moments ago (about the contents of a box; Gopnik & Slaughter, 1991); they can reflect on their own current mental images (Estes, 1994); and they accurately report on their knowledge (or lack thereof) about the contents of a box in front of them (Gonzales, Fabricius, & Kupfer, 2018). Moreover, children who gave such accurate self-reports were more likely to accurately report on *other* people's states of seeing and knowing 7 months later (Gonzales et al., 2018). This form of state self-awareness thus has a scaffolding effect on third-person inferences.

We see that self-awareness, just like other social-cognitive capacities, has layers of complexity: from motor or mind experiences to self-identification to state awareness to memory continuity. Additional levels have barely been researched, such as the emergence of *public* self-awareness (a person's recognition that other people are observing and evaluating the person), which enables the emotions of shame and embarrassment (Chobhthaigh & Wilson, 2015; Lewis, 1997), rich with inferences about the audience's thoughts and evaluations about one's own flawed behavior or character.

Mental State Ascriptions

It should be clear by now that there is no one way to "mentalize"; that many processes connect a perceiver to another's mind. It can be through categorization (e.g., intentionality judgments), attention (e.g., gaze following), coordinated behavior (e.g., imitation), and representation (e.g., inferring knowledge). What is left to discuss are the most sophisticated representations of mental states, demanded by the following challenging circumstances:

- when the states themselves are complex (false belief, self-conscious emotions such as guilt, distinctions such as between jealousy and envy);
- when behavioral evidence for the states is ambiguous (e.g., when a person tries to hide their mental state) or sparse ("what's her goal in sending this email?"); or when the inferred state is a counterfactual ("could she have known?").
- when inferred mental states are combined and incorporated into action explanations ("He was afraid of our reaction and thought that by being quiet we wouldn't notice").
- when the observer wants to know not just *what* another person sees (involved in social referencing) but how another person *interprets* a visual display (e.g., as a "6" or a "9") (see Lalonde & Chandler, 2002).

I call these inferences "ascriptions" to signal that they are often more explicit, with clearer awareness of an "other mind," and are supported by increasingly rich language (e.g., Bartsch & Wellman, 1995) and concepts (Andrews, 2018). Underlying such complex ascriptions are both knowledge-based inferential processes ("he loves hops, he must have a special reason to decline this IPA") and flexible simulation processes ("what would *I* do if I felt so angry?"). Both of them allow the perceiver to go beyond defaults, stereotypical assumptions, and mere projection (Ames, 2004; Clement & Krueger, 2000; Van Boven & Loewenstein, 2003).

Along the developmental path, we are now at the last step of differentiation into a wide range of inferred mental states: not just desires and knowledge, but also false beliefs and intentions. The distinctions emerge fairly gradually and ordered over the course of development from ages 2–7 (Astington, 2001; Flavell, Everett, Croft, & Flavell, 1981; Schult, 2002; Wellman & Liu, 2004) and continue into ages 7–9 if we include third- and fourth-order false beliefs (Osterhaus, Koerber, & Sodian, 2016) as well as action explanations (Atance, Metcalf, Martin-Ordas, & Walker, 2014). Evidence for desire and knowledge inferences in other primates is compelling (e.g., Kaminski, Call, & Tomasello, 2008; Myowa-Yamakoshi, Scola, & Hirata, 2012), but evidence for false-belief inferences is absent (e.g., Call & Tomasello, 2005; Povinelli & Vonk, 2003). Recent studies suggest the possibility that great apes may have an implicit grasp of false beliefs (Buttelmann, Buttelmann, Carpenter, Call, & Tomasello, 2017; Krupenye, Kano, Hirata, Call, & Tomasello, 2016), just as it has been suggested for infants before the age of 2 (Baillargeon, Scott, & He, 2010; Onishi & Baillargeon, 2005; Southgate, Senju, & Csibra, 2007). The interpretation of implicit false-belief results continues to be debated (Andrews, 2018; Perner & Ruffman, 2005; Ruffman & Taumoepeau, 2014), and a number of failed replications of infant results (see Sabbagh & Paulus, 2018) should make us pause and avoid overly strong conclusions. But whatever conclusions we might draw from the implicit tasks, there is little doubt that explicit false-belief ascriptions are robust in 5-year olds and do not occur in 2-year-olds; that great apes fail such explicit false-belief tasks; and that many more explicit mental state inferences are made possible by complex concepts (e.g., emotion categories) and by language (enabling composite representational contents). There is therefore little doubt that social-cognitive capacities ascend in development and evolution and that explicit, contentful mental state ascriptions have evolved and develop late. Consistent with this perspective, we also see that adults take longer to process belief inferences (Malle & Holbrook, 2012; Qureshi, Apperly, & Samson, 2010), have more difficulty at performing them accurately (Epley, Morewedge, & Keysar, 2004; Ickes, 1997; Keysar, 1994), and show the ability for top-down control if there is motivation for improvement (Klein & Hodges, 2001),.

Trait Attributions

"We perceive other people as causal agents, we infer intentions, we infer emotional states, and we go further to infer enduring dispositions or personality traits" (Hastorf, Schneider, & Polefka, 1970). In the spirit of such ascent I placed the process of trait attributions at the top of the tree of social cognition. However, it should not be considered the crowning achievement but rather a consequence of cognitive recombination and abstraction, such that inference of attention, desire, emotion, and belief enable attributions of attitude, temperament, and personality, aided by conceptual distinctions and semantic differentiations. At one point, social psychology treated trait attribution as the most important, frequent, and inevitable tool of social cognition (Jones & Davis, 1965; Shaver, 1975; see Malle, 2011a, for a review); and this trend culminated in the charge that people were "dipositionists" (Ross & Nisbett, 1991), primarily concerned with attributing stable traits or dispositions to others. Against this charge, however, recent evidence shows that people use traits to explain behavior far less often than one would expect (summing to about 5% of explanations; Malle, Knobe, & Nelson, 2007). Moreover, when the behavior in question is highly unusual, some models predicted that trait explanations should increase (Skowronski & Carlston, 1989), but in fact they decrease (Korman & Malle, 2016). Finally, when people encounter text or video displays of ordinary behavior, trait inferences are slower and less prevalent than mental state inferences (Malle & Holbrook, 2012; Van Overwalle et al., 2012).

Along the developmental path, trait attributions seem to emerge later than all other tools of social cognition we have considered. To wit, whether as ascriptions or predictions of future behavior, verbal trait attributions in the good-bad domain begin at age 4 (Boseovski & Lee, 2006). A little later, children make such attributions in the domain of competence: They use evidence for both physical strength and knowledge to make corresponding trait inferences, and those inferences mediate later selective trust to rely on one or another person (Hermes, Behne, & Rakoczy, 2015). Five-year-olds do not yet grasp preferences as traits, instead frequently explaining behavior by reference to norms (Kalish & Shiverick, 2004). Differentiation into trait attributions beyond valence and competence develop from age 6–10 (Gnepp & Chilamkurti, 1988), whereas already 5-year-olds can describe their own personality traits in quite differentiated ways along the Big Five dimensions (Measelle, John, Ablow, Cowan, & Cowan, 2005). This suggests once more that self-directed inferences may facilitate later other-directed inferences.

However, as in many other domains of social-cognitive development, some authors have proposed that infants make trait attributions already at the end of their first year of life. Infants seem to infer that a circle "likes" a triangle that has previously helped the circle (Kuhlmeier, Wynn, & Bloom, 2003), and they prefer agents performing "good" (facilitative) actions over agents performing "bad" (hindering) actions (e.g., Hamlin & Wynn, 2011; Hamlin, Wynn, & Bloom, 2007). Thus, evidence is limited to a proto-moral distinction of good/nice vs. bad/mean. Questions may be raised about such results' specific interpretation and their replicability across different laboratories (Margoni & Surian, 2018), but it is a plausible hypothesis that implicit trait attributions along the valence dimension launch the ability to more generally attribute traits to others. In their full-fledged form, trait attributions occur along a host of dimensions (not just valence), come in degrees (not just categories), and exist within a conceptual space that includes temperament, personality, moral character, values, and ideology. Arriving at this sophisticated space of trait attributions requires a good deal of concepts and language, experiences with a variety of individuals, and an understanding of the stability but also context specificity of traits.

The Tree, Once More

Some of the evidence I have reported on the likely development of social-cognitive capacities and on their presence in nonhuman primates is incomplete, open to interpretation, or still under debate. Nonetheless, Fig. 2 offers a tentative summary of evidence on the developmental time scale and an even more tentative assessment of evidence from the animal behavior literature. Within the latter, brighter shades of gray indicate higher confidence for a capacity's presence in nonhuman animals in light of scholarly consensus on replicated evidence, both in field and lab; darker shades of gray indicate lower confidence in light of scholarly consensus on a capacity's absence or simply absence of evidence. In between are mixed data and debate.

One repeated theme in the overview of these social-cognitive capacities is that the capacities vary in their degree of representing the actual *mind* of another (not just their behavior) and how many knowledge structures aide this representation. In addition, many of the capacities themselves come in such degrees of mind representation (e.g., variants of gaze following, social attention, imitation), and mentalizing may therefore be seen as a bundle of continua. A second repeated theme is the impact of self-awareness, self-regulation, and of social context in differentiating and modifying numerous capacities, thus providing important functions that make social living possible. A third theme is the affinity and facilitation among many of the capacities; I now elaborate on these relations.

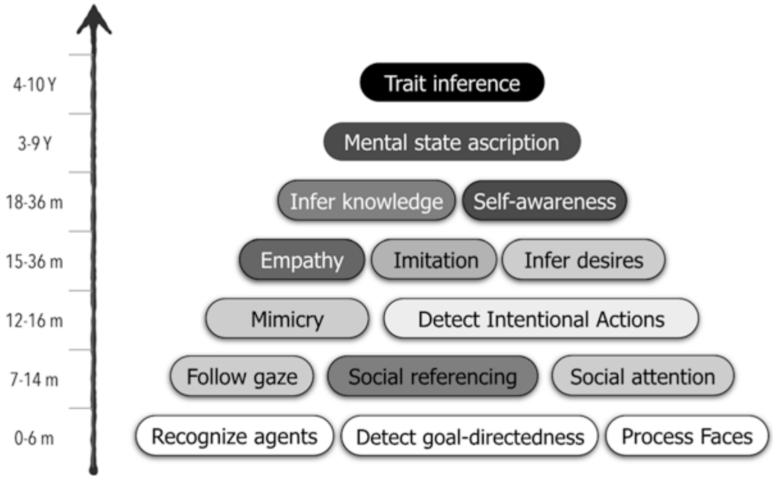


Fig. 2 The tree of social cognition and its hierarchically ordered processes, roughly aligned with a time scale of emergence in human development and shaded by likely presence, given current evidence, in nonhuman animals (the darker the shading the less likely to be present)

Hierarchical Dependencies

Perhaps the most important feature of a hierarchical conception of social cognition is that the results of many lower-order tools (often in combination) are inputs to the processing performed by higher-order tools. This characterization has affinity with models of hierarchical cognitive control (Badre & Nee, 2018) but runs counter to the picture of a dichotomous division into two levels or systems of mentalizing (e.g., Apperly & Butterfill, 2009; Coricelli, 2005), akin to the well-known "System 1/System 2" division. Though it is likely that, in general, lower-order capacities (LC) tend to be "automatic" and "unconscious" and higher-order capacities (HC) tend to be "reflective" and "conscious," such assignments should not be considered categorical or fixed. LC can become reflective (e.g., an intentionality judgment in the jury box), and HC can become automatized in the presence of familiar stimuli (e.g., repeated inferences of specific mental states for close others). In the picture of a tree of social cognition, some bundles of LC tend to be engaged first (and perhaps continuously) to solve certain social challenges, and some HC step in to integrate this early information or take over when the LC cannot by themselves solve the challenge at hand. For example, LC may track referents in a conversation and HC may try to resolve a possible misunderstanding in the conversation. Many such LC-HC relationships exist, and I describe three of them below. I indicate joint operation with a "+" sign and facilitation with the notation "⇉."

Recognizing Agent + Detecting Goal-Directedness + Gaze Following ⇉ Detecting Intentionality + Infer Desires

Once people identify agents as the entities of greatest interest to them—e.g., by noticing eyes or experiencing contingent responses—they can appropriately code actions as goal-directed toward certain objects. By attending to breakpoints in the behavior stream (e.g., turning body and head, movement slowing just before object touch) and tracking gaze as well as selective physical actions (e.g., grasping one rather than another object), the perceiver can recognize the equifinality of a behavior and, with additional observation of the object, infer the agent's likely desire (e.g., food is desired for eating, complex objects are desired for taking apart). With repetition, certain movement patterns (e.g., shaking hands, putting down keys) become distinct configurations and are instantaneously recognized as "intentional."

Process Faces + Social Referencing + Mimicry ⇉ Empathy

With improved decoding of facial expressions comes a more refined capacity for mimicry, which has long been considered the basis for emotional contagion (Hatfield et al., 1994) and empathy (Lipps, 1907). But even though people mimic others' emotional expressions and are able to empathize with others, the mimicking itself may rarely cause the empathic response directly (Hess & Fischer, 2013). Nonetheless, mimicry may be an indicator of a dispositionally heightened responsiveness to others' behavior (Franzen, Mader, & Winter, 2018; Sonnby-Borgström, Jönsson, & Svensson, 2003), and if mimicry is reciprocal, it can contribute indirectly to emotion matching by stabilizing each person's emotion (and mutual empathy) through stabilizing their expression.

Furthermore, with improved decoding of facial expressions and body postures come more opportunities for social referencing. To the extent that this referencing process often aligns people's evaluations, it will also align their emotions (and expressions thereof, which could look like mimicry). Such emotion matching is not a form of contagion but arises from recognizing how the other evaluates an object, action, or person and adopting (or agreeing with) this evaluation. Finally, simulation of others' feelings and ascriptions of specific emotions can create congruent emotional expressions (Hawk et al., 2011).

Social Attention + Detecting Intentionality + Infer Desires ⇉ Imitation

Mature joint attention and social referencing processes allow agents to align their attention and evaluations for shared experience and joint actions, including both complementary and imitative behavior, suggesting a facilitative linkage between attention and imitation (Kana, Wadsworth, & Travers, 2011). Recognizing the other's intentional actions and object-specific desires further facilitates imitation, because the perceiver understands not only the other's observable behavior but their "invisible" goals.

Postscriptum: Concepts

I have not said much about concepts, even though I am on record for proposing that "theory of mind" is first and foremost a conceptual framework (Malle, 2005, 2008). My current view is that several of the described processes of social cognition build on basic categorizations (e.g., into agents and nonagents, intentional vs. unintentional behaviors) that are initially aided by sensitivity to certain perceptual markers and reinforced in social interaction. Over the course of development, many processes of social cognition get more refined and build up abstractions that form finer-grained concepts, such as shades of desire, intention, belief, knowledge, etc. All the basic and more fine-grained concepts guide information search and processing—such as when the agent category initiates gaze following or when the intentionality category triggers a desire inference. Moreover, as these concepts mature, they can stand in specific (again, often hierarchical) logical relations to one another and shape expectations about what can actually be observed. For instance, observing an intentional action implies that the agent had some desire and belief, and further inferential processes have to determine what those states are. The intentionality concept, in particular, grows into a complex but systematic conceptual structure (Malle & Knobe, 1997, 2001) whose embedded processes guide both moral judgments (Monroe & Malle, 2017) and behavior explanations (Malle, 2011b).

Climb to the Top: The Cultural History of Mental State Ascriptions

Studying this broad literature has convinced me that the emergence of mentalizing as full-blown mental state ascriptions, and of trait ascriptions building on them, requires many steps: maturation, continued learning, social interactions that scaffold, and reliance on lower-order capacities. Language, moreover, facilities conceptual distinctions (e.g., Barsalou, 1983) and is therefore a key vehicle to support these ascriptions and their increasing differentiations—for example, among shades of desires and intention (wanting, planning, intending, deciding, committing; Malle & Knobe, 2001) or among the many shades of emotions. I want to put forth an additional hypothesis, not about the emergence of the *principled* ability to make mental state ascriptions, but about a powerful increase in the use and significance of such ascriptions in human cultural evolution: the hypothesis that mental state ascriptions, and also trait ascriptions, exploded after humans settled down about 12,000 years ago.

Sedentism was caused by and caused a considerable number of cascading changes: population increase, agriculture and animal husbandry, religion, architecture, organized fighting, and many more (Aurenche, Kozłowski, & Kozłowski, 2013; Boserup, 1965; Peregrine, Ember, & Ember, 2007; Redman, 1978; Renfrew, 2007; Zeder, 2011). I will focus on the ones that may have specifically contributed to the rise of mental state ascriptions.[2]

Population Explosion

Between 10,000 and 8000 BCE, a first population growth began in many human settlements from camps to villages and towns across Europe and West Asia (Atkinson, Gray, & Drummond, 2008; Gignoux, Henn, & Mountain, 2011; Hawks, Wang, Cochran, Harpending, & Moyzis, 2007; Lee, 1972). Among the causes of this growth were broader environmental opportunities (the end of the Last Glacial;

[2] For the sake of a cultural history perspective, I express the reported background evidence as claims about the past, even when actual archeological evidence is often lacking and our knowledge stems from the study of present-time hunter-gatherer and sedentary societies.

Shultziner et al., 2010), specific local circumstances of fauna and flora (Aurenche et al., 2013), but also the impact of settlement on child bearing. In nomadic communities, mothers had to carry their newborns for thousands of kilometers a year and breast-fed for a longer time, which limited them to one child every 3–4 years (Lee, 1972; Shostak, 2009). Once settled, pregnancy frequencies increased substantially, dramatically raising child birth rates (Buikstra, Konigsberg, & Bullington, 1986).

As families grew in size, kinship became a stricter boundary of ingroup and outgroup (Alt et al., 2013; Wilson, 1988). Thus, empathy and imitation were practiced more within families than in the community at large. In contrast to living in hunter-gatherer groups of 10–100 (Williams, 1985), in which most everybody knew everybody else, living in communities increasing to 1000 (around 7000 BCE) and later to 100,000 (around 1200 BCE; Modelski, 2003) created significant relational distance (Sutcliffe, Dunbar, Binder, & Arrow, 2012). The sheer number of people, and especially the number of people with whom one had weak or no relations, made understanding more difficult, owing to fewer interactions, fewer joint experiences, and more suspicion about the other's benevolent motives. To overcome such gaps of understanding, uncertainty, and threats of conflict, mental state ascriptions must have gained in importance.

Visibility

As towns grew into cities and empires, built structures rapidly increased in number, size, and complexity (Flannery, 2002; Wilson, 1988). Buildings created barriers, defined spheres of inside and outside, private and public (Duru, 2018; Hodder, 1990). When actions and minds were hidden behind private walls, people could no longer attend to and monitor each other (Wilson, 1988) and needed to exert additional efforts to recognize others as understandable and trustworthy. None of the lower-order capacities we have examined operate at a distance; only the two highest-order ones do. Given the increased use of mind perception at a distance, we can also better understand the rise of agentic, doctrinal religion (Cauvin, 2000; Dunbar, 2013; Hodder, 2018), in which the minds of Gods and spirits were objects of heavy mental state ascriptions (Guthrie, 1993; Tremlin, 2006), and are so to this day (Heiphetz, Lane, Waytz, & Young, 2016).

Diversification

By staying in place, people had more opportunity and benefit for differentiation of practices, crafts, and positions in society (Benz & Bauer, 2013). This diversity demanded tracking of different agents' motives and traits and updating that knowledge in each interaction. Moreover, the explosion of tools manufactured for food production and building construction required many different skills and, aided by genetic diversification (Ricaut et al., 2012), individual differences in abilities and personality increased. This in turn commanded complex trait inferences and their underlying mental state inferences.

Possessions, Law, and War

Over the millennia, ownership of land, livestock, and tools led to wealth that was inherited within families, thus further intensifying kinship boundaries. Accumulated wealth came with threats to lose it and with competition for more wealth through vending and trading. This situation called for norms and laws of inheritance, theft, and economic exchange (Binder, 2002; Milgrom, North, & Weingast, 1990), along with institutional forms of enforcement and accompanying requirements for mental state ascriptions to keep such enforcements fair (Monroe & Malle, 2019; Voiklis & Malle, 2017). The law, of course, famously implements many of the fundamental distinctions of mental state ascription (Duff, 1990; Marshall, 1968). At a societal level, territorial expansions provoked broader and more frequent intergroup conflict that gave rise to organized warfare. Whereas a duel of two individuals can rely on many of the lower-order social-cognitive capacities in shared space, organized warfare is collective, tactical, strategic, and thus requires social cognition at a distance, leaving once more only the highest-order capacities in contention.

Implications and Open Questions

This is then my picture of mentalizing: a broad, closely intertwined hierarchy of social-cognitive capacities, among which the late-developing, slower, and cognitively demanding forms were substantially amplified in very recent human history. This picture offers a number of implications and unanswered questions, three of which I touch on; and it demands numerous revisions, which the scientific community at large, I hope, will undertake.

Theoretical Pluralism

The tree of social cognition welcomes a diversity of theoretical positions and structures: simulation theory's self-based models, theory theory's inference processes, massive bottom-up learning, abstract concepts, and even various degrees of "preparedness." It seems doubtful that any of the branches of this tree are completely encapsulated processes (Fodor, 1983); but sensitivities to certain stimuli (e.g., biological movement, eyes) may indeed be formative in the human mind with little learning, and mimicry may be facilitated by old and ready mappings between visual representations and motor programs (Iacoboni, 2009). Most other capacities, however, are complex and rely on multiple interacting processes, grow with experience, and benefit from social scaffolding (Barrett, 2015). The tree even provides space for the somewhat radical claims of scholars who question whether others' minds are "hidden" (Gallagher, 2008; Gibbs, 1999; Hutto, 2007). There is truth in the claim that often we are not "thinking about what might be going on in the other person's mind" (Gallagher, 2008, p. 540); the numerous lower-order social-cognitive processes certainly attest to that. But we also must acknowledge the substantial role of higher-order, explicit mentalizing, especially after individual development and practice, and in the vast society of strangers *homo sapiens* has formed.

Measurement

If social cognition is conceptualized as a hierarchical network of more than a dozen processes, their distinct measurement is a major challenge, especially if we want to put the claims of hierarchical and facilitative relationships to a test. Developmental and comparative psychologists have done impressive work in creating and collating such measures and experimental tasks for infants, children, and animals (see Herrmann et al., 2007, for a particularly commendable project). For assessments in adulthood, tests have been designed in different literatures and, because of their separation, have provided very little information about discriminative validity. A review and evaluation of these literatures goes beyond this chapter but would obviously be worthwhile. Once such measures have been validated behaviorally and cognitively, we would be able to systematically examine cultural variations, neural correlates, or genetic markers.

Cultural Factors

The hypothesis of recent cultural pressures on the practice and refinement of mental state and trait ascriptions poses interesting challenges and suggestions. In particular, it encourages expansion of existing lines of research on the differential engagement of social cognition in remote settings vs. co-presence; for strangers vs. close others; for ingroup vs. outgroup members; or in competitive vs. cooperative contexts (e.g., Ames, 2004; Haslam, 2006; Lin, Qu, & Telzer, 2018). Cross-cultural variations may also be studied in a more nuanced matter—less as a categorical difference between East and West and more as a function of the differential learning and the social-cognitive challenges that come with demands, tasks, and rewards that particular cultural contexts provide.

References

Alicke, M. D., Buckingham, J., Zell, E., & Davis, T. (2008). Culpable control and counterfactual reasoning in the psychology of blame. *Personality and Social Psychology Bulletin, 34*, 1371–1381. https://doi.org/10.1177/0146167208321594

Alt, K. W., Benz, M., Mueller, W., Berner, M. E., Schultz, M., Schmidt-Schultz, T. H., et al. (2013). Earliest evidence for social endogamy in the 9,000-year-old-population of Basta, Jordan. *PLoS One, 8*, e65649. https://doi.org/10.1371/journal.pone.0065649

Ames, D. R. (2004). Inside the mind reader's tool kit: Projection and stereotyping in mental state inference. *Journal of Personality and Social Psychology, 87*, 340–353. https://doi.org/10.1037/0022-3514.87.3.340

Andrews, K. (2018). Apes track false beliefs but might not understand them. *Learning & Behavior, 46*, 3–4. https://doi.org/10.3758/s13420-017-0288-8

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*, 953–970. https://doi.org/10.1037/a0016923

Astington, J. W. (2001). The paradox of intention: Assessing children's metarepresentational understanding. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 85–103). Cambridge, MA: MIT Press.

Atance, C. M., Metcalf, J. L., Martin-Ordas, G., & Walker, C. L. (2014). Young children's causal explanations are biased by post-action associative information. *Developmental Psychology, 50*, 2675–2685. https://doi.org/10.1037/a0038186

Atance, C. M., & O'Neill, D. K. (2004). Acting and planning on the basis of a false belief: Its effects on 3-year-old children's reasoning about their own false beliefs. *Developmental Psychology, 40*, 953–964. https://doi.org/10.1037/0012-1649.40.6.953

Atkinson, Q. D., Gray, R. D., & Drummond, A. J. (2008). mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Molecular Biology and Evolution, 25*, 468–474. https://doi.org/10.1093/molbev/msm277

Aurenche, O., Kozłowski, J. K., & Kozłowski, S. K. (2013). To be or not to be … Neolithic: "Failed attempts" at Neolithization in Central and Eastern Europe and in the Near East, and their final success (35,000-7000 Bp). *Paléorient, 39*, 5–45.

Badre, D., & Nee, D. E. (2018). Frontal cortex and the hierarchical control of behavior. *Trends in Cognitive Sciences, 22*, 170–188. https://doi.org/10.1016/j.tics.2017.11.005

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14*, 110–118. https://doi.org/10.1016/j.tics.2009.12.006

Bakeman, R., & Adamson, L. B. (1984). Coordinating attention to people and objects in mother–infant and peer–infant interaction. *Child Development, 55*, 1278–1289. https://doi.org/10.2307/1129997

Baldwin, D. A., & Baird, J. A. (2001). Discerning intentions in dynamic human action. *Trends in Cognitive Sciences, 5*, 171–178. https://doi.org/10.1016/S1364-6613(00)01615-6

Barker, M., & Givón, T. (2005). Representation of the interlocutor's mind during conversation. In B. F. Malle & S. D. Hodges (Eds.), *Other minds: How humans bridge the divide between self and others* (pp. 223–238). New York, NY: Guilford Press.

Barragan-Jason, G., Cauchoix, M., & Barbeau, E. J. (2015). The neural speed of familiar face recognition. *Neuropsychologia, 75*, 390–401. https://doi.org/10.1016/j.neuropsychologia.2015.06.017

Barrett, H. C. (2015). *The shape of thought: How mental adaptations evolve*. Oxford: Oxford University Press.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition, 11*, 211–227. https://doi.org/10.3758/BF03196968

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. New York, NY: Oxford University Press.

Begus, K., & Southgate, V. (2012). Infant pointing serves an interrogative function. *Developmental Science, 15*, 611–617. https://doi.org/10.1111/j.1467-7687.2012.01160.x

Benz, M., & Bauer, J. (2013). Symbols of power – Symbols of crisis? A psycho-social approach to early neolithic symbol systems. *Neo-Lithics: The Newsletter of Southwest Asian Neolithic Research, 2*(2013), 11–24.

Berry, D. S., Kean, K. J., Misovich, S. J., & Baron, R. M. (1991). Quantized displays of human movement: A methodological alternative to the point-light display. *Journal of Nonverbal Behavior, 15*, 81–97. https://doi.org/10.1007/BF00998264

Binder, G. (2002). Punishment theory: Moral or political? *Buffalo Criminal Law Review, 5*, 321–372.

Blairy, S., Herrera, P., & Hess, U. (1999). Mimicry and the judgment of emotional facial expressions. *Journal of Nonverbal Behavior, 23*, 5–41. https://doi.org/10.1023/A:1021370825283

Blocker, H. S., & McIntosh, D. N. (2016). Automaticity of the interpersonal attitude effect on facial mimicry: It takes effort to smile at neutral others but not those we like. *Motivation and Emotion, 40*, 914–922. https://doi.org/10.1007/s11031-016-9581-7

Bloom, P. (2007). Religion is natural. *Developmental Science, 10*, 147–151. https://doi.org/10.1111/j.1467-7687.2007.00577.x

Boseovski, J. J., & Lee, K. (2006). Children's use of frequency information for trait categorization and behavioral prediction. *Developmental Psychology, 42*, 500–513. https://doi.org/10.1037/0012-1649.42.3.500

Boserup, E. (1965). *The conditions of agricultural growth: The economics of agrarian change under population pressure*. New York, NY: Aldine.

Brooks, R., & Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science, 8*, 535–543. https://doi.org/10.1111/j.1467-7687.2005.00445.x

Buikstra, J. E., Konigsberg, L. W., & Bullington, J. (1986). Fertility and the development of agriculture in the prehistoric Midwest. *American Antiquity, 51*, 528–546. https://doi.org/10.2307/281750

Buttelmann, D., Buttelmann, F., Carpenter, M., Call, J., & Tomasello, M. (2017). Great apes distinguish true from false beliefs in an interactive helping task. *PLoS One, 12*, e0173793. https://doi.org/10.1371/journal.pone.0173793

Call, J., & Tomasello, M. (1998). Distinguishing intentional from accidental actions in orangutans (Pongo pygmaeus), chimpanzees (Pan troglodytes) and human children (Homo sapiens). *Journal of Comparative Psychology, 112*, 192–206. https://doi.org/10.1037/0735-7036.112.2.192

Call, J., & Tomasello, M. (2005). What chimpanzees know about seeing, revisited: An explanation of the third kind. In *Joint attention: Communication and other minds: Issues in philosophy and psychology, Consciousness and self-consciousness* (pp. 45–64). New York, NY: Clarendon Press/Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199245635.003.0003

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences, 12*, 187–192. https://doi.org/10.1016/j.tics.2008.02.010

Cameron, C. D., Hutcherson, C. A., Ferguson, A. M., Scheffer, J. A., Hadjiandreou, E., & Inzlicht, M. (2019). Empathy is hard work: People choose to avoid empathy because of its cognitive costs. *Journal of Experimental Psychology: General, 148*, 962–976. https://doi.org/10.1037/xge0000595

Campbell, M. W., & de Waal, F. B. M. (2014). Chimpanzees empathize with group mates and humans, but not with baboons or unfamiliar chimpanzees. *Proceedings of the Royal Society B: Biological Sciences, 281*, 20140013. https://doi.org/10.1098/rspb.2014.0013

Carpenter, M., Akhtar, N., & Tomasello, M. (1998). Fourteen- through 18-month-old infants differentially imitate intentional and accidental actions. *Infant Behavior and Development, 21*, 315–330. 16/S0163-6383(98)90009-1 [pii].

Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development, 63*, 176–176. https://doi.org/10.2307/1166214

Carrasco, L., Posada, S., & Colell, M. (2009). New evidence on imitation in an enculturated chimpanzee (Pan troglodytes). *Journal of Comparative Psychology, 123*, 385–390. https://doi.org/10.1037/a0016275

Cauvin, J. (2000). *The birth of the gods and the origins of agriculture*. Cambridge: Cambridge University Press.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology, 76*, 893–910. https://doi.org/10.1037/0022-3514.76.6.893

Cheung, E. O., Slotter, E. B., & Gardner, W. L. (2015). Are you feeling what I'm feeling? The role of facial mimicry in facilitating reconnection following social exclusion. *Motivation and Emotion, 39*, 613–630. https://doi.org/10.1007/s11031-015-9479-9

Chobhthaigh, S. N., & Wilson, C. (2015). Children's understanding of embarrassment: Integrating mental time travel and mental state information. *British Journal of Developmental Psychology, 33*, 324–339. https://doi.org/10.1111/bjdp.12094

Chouinard, M. M. (2007). Children's questions: A mechanism for cognitive development. *Monographs of the Society for Research in Child Development, 72*, 1–129.

Clark, H. H. (1996). *Using language*. New York, NY: Cambridge University Press.

Clay, Z., & Tennie, C. (2018). Is overimitation a uniquely human phenomenon? Insights from human children as compared to bonobos. *Child Development, 89*, 1535–1544. https://doi.org/10.1111/cdev.12857

Clement, R. W., & Krueger, J. (2000). The primacy of self-referent information in perceptions of social consensus. *British Journal of Social Psychology, 39*, 279–299. https://doi.org/10.1348/014466600164471

Coricelli, G. (2005). Two-levels of mental states attribution: From automaticity to voluntariness. *Neuropsychologia, Movement, Action and Consciousness: Toward a Physiology of Intentionality. A Special Issue in Honour of Marc Jeannerod, 43*, 294–300. https://doi.org/10.1016/j.neuropsychologia.2004.11.015

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*, 353–380. https://doi.org/10.1016/j.cognition.2008.03.006

de Haan, M., & Nelson, C. A. (1999). Brain activity differentiates face and object processing in 6-month-old infants. *Developmental Psychology, 35*, 1113–1121. https://doi.org/10.1037/0012-1649.35.4.1113

de Waal, F. B. M., & Preston, S. D. (2017). Mammalian empathy: Behavioural manifestations and neural basis. *Nature Reviews Neuroscience, 18*, 498–509. https://doi.org/10.1038/nrn.2017.72

Decety, J., & Cacioppo, S. (2012). The speed of morality: A high-density electrical neuroimaging study. *Journal of Neurophysiology, 108*, 3068–3072. https://doi.org/10.1152/jn.00473.2012

Di Giorgio, E., Turati, C., Altoè, G., & Simion, F. (2012). Face detection in complex visual displays: An eye-tracking study with 3- and 6-month-old infants and adults. *Journal of Experimental Child Psychology, 113*, 66–77. https://doi.org/10.1016/j.jecp.2012.04.012

Dodonova, Y. A., & Dodonov, Y. S. (2012). Speed of emotional information processing and emotional intelligence. *International Journal of Psychology, 47*, 429–437. https://doi.org/10.1080/00207594.2012.656131

Duff, R. A. (1990). *Intention, agency and criminal liability*. Oxford: Basil Blackwell.

Dunbar, R. I. M. (2013). What makes the neolithic so special. *Neo-Lithics: The Newsletter of Southwest Asian Neolithic Research, 2*(2013), 25–29.

Dunbar, R. I. M., & Shultz, S. (2007). Evolution in the social brain. *Science, 317*, 1344–1347. https://doi.org/10.1126/science.1145463

Duru, G. (2018). Sedentism and solitude: Exploring the impact of private space on social cohesion in the neolithic. In I. Hodder (Ed.), *Religion, history, and place in the origin of settled life* (pp. 162–185). Louisville, CO: University Press of Colorado. https://doi.org/10.2307/j.ctv3c0thf.11

Eckerman, C. O., Davis, C. C., & Didow, S. M. (1989). Toddlers' emerging ways of achieving social coordinations with a peer. *Child Development, 60*, 440–453. https://doi.org/10.2307/1130988

Epley, N., Morewedge, C. K., & Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology, 40*, 760–768. https://doi.org/10.1016/j.jesp.2004.02.002

Estes, D. (1994). Young children's understanding of the mind: Imagery, introspection, and some implications. *Journal of Applied Developmental Psychology, 15*, 529–548. https://doi.org/10.1016/0193-3973(94)90021-3

Feinman, S. (1992). Social referencing and conformity. In S. Feinman (Ed.), *Social referencing and the social construction of reality in infancy* (pp. 229–267). Boston, MA: Springer. https://doi.org/10.1007/978-1-4899-2462-9_10

Feshbach, N. D., & Roe, K. (1968). Empathy in six- and seven-year-olds. *Child Development, 39*, 133–145. https://doi.org/10.2307/1127365

Fiala, B., Arico, A., & Nichols, S. (2014). You, robot. In E. Machery & E. O'Neill (Eds.), *Current controversies in experimental philosophy* (pp. 31–47). London: Routledge.

Fischer, A., & Hess, U. (2017). Mimicking emotions. *Current Opinion in Psychology, Emotion, 17*, 151–155. https://doi.org/10.1016/j.copsyc.2017.07.008

Flannery, K. V. (2002). The origins of the village revisited: From nuclear to extended households. *American Antiquity, 67*, 417–433. https://doi.org/10.2307/1593820

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. *Developmental Psychology, 17*, 99–103. https://doi.org/10.1037/0012-1649.17.1.99

Flavell, J. H., Green, F. L., & Flavell, E. R. (1995). Young children's knowledge about thinking. *Monographs of the Society for Research in Child Development, 60*, v–96. https://doi.org/10.2307/1166124

Flynn, E., & Smith, K. (2012). Investigating the mechanisms of cultural acquisition: How pervasive is overimitation in adults? *Social Psychology, 43*, 185–195. https://doi.org/10.1027/1864-9335/a000119

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Franco, F. (2005). Infant pointing: Harlequin, servant of two masters. In N. Eilan, C. Hoerl, T. McCormack, & J. Roessler (Eds.), *Joint attention: Communication and other minds: Issues in philosophy and psychology* (pp. 129–164). New York, NY: Clarendon Press/Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199245635.003.0007

Franzen, A., Mader, S., & Winter, F. (2018). Contagious yawning, empathy, and their relation to prosocial behavior. *Journal of Experimental Psychology: General, 147*, 1950–1958. https://doi.org/10.1037/xge0000422

Fukumura, K. (2015). Interface of linguistic and visual information during audience design. *Cognitive Science, 39*, 1419–1433. https://doi.org/10.1111/cogs.12207

Fusaro, M., & Harris, P. L. (2013). Dax gets the nod: Toddlers detect and use social cues to evaluate testimony. *Developmental Psychology, Selective Social Learning, 49*, 514–522. https://doi.org/10.1037/a0030580

Gallagher, S. (2008). Direct perception in the intersubjective context. *Consciousness and Cognition: An International Journal, 17*, 535–543. https://doi.org/10.1016/j.concog.2008.03.003

Gallup, G. G. (1970). Chimpanzees: Self-recognition. *Science, 167*, 86–87. https://doi.org/10.1126/science.167.3914.86

Gibbs, R. W., Jr. (1999). *Intentions in the experience of meaning*. New York, NY: Cambridge University Press.

Gignoux, C. R., Henn, B. M., & Mountain, J. L. (2011). Rapid, global demographic expansions after the origins of agriculture. *Proceedings of the National Academy of Sciences, 108*, 6044–6049. https://doi.org/10.1073/pnas.0914274108

Gnepp, J., & Chilamkurti, C. (1988). Children's use of personality attributions to predict other people's emotional and behavioral reactions. *Child Development, 59*, 743–754. https://doi.org/10.2307/1130573

Goldman, A. (2009). Précis of simulating minds: The philosophy, psychology, and neuroscience of mindreading. *Philosophical Studies, 144*, 431–434. https://doi.org/10.1007/s11098-009-9355-0

Gonzales, C. R., Fabricius, W. V., & Kupfer, A. S. (2018). Introspection plays an early role in children's explicit theory of mind development. *Child Development, 89*, 1545–1552. https://doi.org/10.1111/cdev.12876

Gopnik, A., & Slaughter, V. (1991). Young children's understanding of changes in their mental states. *Child Development, 62*, 98–110. https://doi.org/10.2307/1130707

Gordon, R. (1986). Folk psychology as simulation. *Mind & Language, 1*, 158–171. https://doi.org/10.1111/j.1468-0017.1986.tb00324.x

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*, 125–130. https://doi.org/10.1016/j.cognition.2012.06.007

Gredebäck, G., Astor, K., & Fawcett, C. (2018). Gaze following is not dependent on ostensive cues: A critical test of natural pedagogy. *Child Development, 89*, 2091–2098. https://doi.org/10.1111/cdev.13026

Gruber, T., Deschenaux, A., Frick, A., & Clément, F. (2019). Group membership influences more social identification than social learning or overimitation in children. *Child Development, 90*, 728. https://doi.org/10.1111/cdev.12931

Guthrie, S. E. (1993). *Faces in the clouds: A new theory of religion*. New York, NY: Oxford University Press.

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development, 26*, 30–39. https://doi.org/10.1016/j.cogdev.2010.09.001

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*, 557–559. https://doi.org/10.1038/nature06288

Harris, P. L., Yang, B., & Cui, Y. (2017). 'I don't know': Children's early talk about knowledge. *Mind & Language, 32*, 283–307. https://doi.org/10.1111/mila.12143

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc, 10*, 252–264. https://doi.org/10.1207/s15327957pspr1003_4

Hastorf, A. H., Schneider, D. J., & Polefka, J. (1970). *Person perception*. Reading, MA: Addison-Wesley.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1994). *Emotional contagion*. New York, NY: Cambridge University Press.

Hawk, S. T., Fischer, A. H., & Van Kleef, G. A. (2011). Taking your place or matching your face: Two paths to empathic embarrassment. *Emotion, 11*, 502–513. https://doi.org/10.1037/a0022762

Hawks, J., Wang, E. T., Cochran, G. M., Harpending, H. C., & Moyzis, R. K. (2007). Recent acceleration of human adaptive evolution. *Proceedings of the National Academy of Sciences, 104*, 20753–20758. https://doi.org/10.1073/pnas.0707650104

Heider, F. (1958). *The psychology of interpersonal relations*. New York, NY: Wiley.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology, 57*, 243–259.

Heiphetz, L., Lane, J. D., Waytz, A., & Young, L. L. (2016). How children and adults represent god's mind. *Cognitive Science, 40*, 121–144. https://doi.org/10.1111/cogs.12232

Hermes, J., Behne, T., & Rakoczy, H. (2015). The role of trait reasoning in young children's selective trust. *Developmental Psychology, 51*, 1574–1587. https://doi.org/10.1037/dev0000042

Herrmann, E., Call, J., Hernàndez-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science (New York, N.Y.), 317*, 1360–1366. https://doi.org/10.1126/science.1146282

Hess, U., & Blairy, S. (2001). Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy. *International Journal of Psychophysiology, 40*, 129–141. https://doi.org/10.1016/S0167-8760(00)00161-6

Hess, U., & Fischer, A. (2013). Emotional mimicry as social regulation. *Personality and Social Psychology Review, 17*, 142–157. https://doi.org/10.1177/1088868312472607

Hess, U., Houde, S., & Fischer, A. (2014). *Do we mimic what we see or what we know?* Oxford: Oxford University Press.

Hewlett, B. S., Berl, R. E. W., & Roulette, C. J. (2016). Teaching and overimitation among Aka hunter-gatherers. In H. Terashima & B. S. Hewlett (Eds.), *Social learning and innovation in contemporary hunter-gatherers: Evolutionary and ethnographic perspectives* (pp. 35–45). New York, NY: Springer.

Heyes, C. (2011). Automatic imitation. *Psychological Bulletin, 137*, 463–483. https://doi.org/10.1037/a0022288

Higgins, E. T., & Pittman, T. S. (2008). Motives of the human animal: Comprehending, managing, and sharing inner states. *Annual Review of Psychology, 59*, 361–385. https://doi.org/10.1146/annurev.psych.59.103006.093726

Hodder, I. (1990). *The domestication of Europe*. Oxford: Basil Blackwell.

Hodder, I. (Ed.). (2018). *Religion, history, and place in the origin of settled life*. Louisville, CO: University Press of Colorado.

Hodges, S. D., & Klein, K. J. K. (2001). Regulating the costs of empathy: The price of being human. *The Journal of Socio-Economics, 30*, 437–452. https://doi.org/10.1016/S1053-5357(01)00112-3

Hoerl, C., & McCormack, T. (2005). Joint reminiscing as joint attention to the past. In N. Eilan, C. Hoerl, T. McCormack, & J. Roessler (Eds.), *Joint attention: Communication and other minds: Issues in philosophy and psychology* (pp. 260–286). New York, NY: Clarendon Press/Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199245635.003.0012

Huang, C.-T., Heyes, C., & Charman, T. (2006). Preschoolers' behavioural reenactment of "failed attempts": The roles of intention-reading, emulation and mimicry. *Cognitive Development, 21*, 36–45. https://doi.org/10.1016/j.cogdev.2005.09.002

Hutto, D. D. (2007). The narrative practice hypothesis: Origins and applications of folk psychology. *Royal Institute of Philosophy Supplement, 60*, 43–68. https://doi.org/10.1017/S1358246107000033

Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annual Review of Psychology, 60*, 653–670. https://doi.org/10.1146/annurev.psych.60.110707.163604

Ichikawa, H., Kanazawa, S., Yamaguchi, M. K., & Kakigi, R. (2010). Infant brain activity while viewing facial movement of point-light displays as measured by near-infrared spectroscopy (NIRS). *Neuroscience Letters, 482*, 90–94. https://doi.org/10.1016/j.neulet.2010.06.086

Ickes, W. (Ed.). (1997). *Empathic accuracy*. New York, NY: Guilford Press.

Jessen, S., & Grossmann, T. (2015). Neural signatures of conscious and unconscious emotional face processing in human infants. *Cortex, 64*, 260–270. https://doi.org/10.1016/j.cortex.2014.11.007

Johnson, K. L., & Shiffrar, M. (2013). *People watching: Social, perceptual, and neurophysiological studies of body perception. Oxford series in visual cognition*. New York, NY: Oxford University Press.

Johnson, S. C., Booth, A., & O'Hearn, K. (2001). Inferring the goals of a nonhuman agent. *Cognitive Development, 16*, 637–656. https://doi.org/10.1016/S0885-2014(01)00043-0

Johnson, S. C., Shimizu, Y. A., & Ok, S.-J. (2007). Actors and actions: The role of agent behavior in infants' attribution of goals. *Cognitive Development, 22*, 310–322. https://doi.org/10.1016/j.cogdev.2007.01.002

Johnson, S. C., Slaughter, V., & Carey, S. (1998). Whose gaze will infants follow? The elicitation of gaze-following in 12-month-olds. *Developmental Science, 1*, 233–238. https://doi.org/10.1111/1467-7687.00036

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 219–266). New York, NY: Academic Press.

Jones, S. S. (2007). Imitation in infancy: The development of mimicry. *Psychological Science, 18*, 593–599. https://doi.org/10.1111/j.1467-9280.2007.01945.x

Jones, S. S. (2009). The development of imitation in infancy. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 364*, 2325–2335. https://doi.org/10.1098/rstb.2009.0045

Kalish, C. W., & Shiverick, S. M. (2004). Children's reasoning about norms and traits as motives for behavior. *Cognitive Development, 19*, 401–416.

Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition, 109*, 224–234. https://doi.org/10.1016/j.cognition.2008.08.010

Kana, R. K., Wadsworth, H. M., & Travers, B. G. (2011). A systems level analysis of the mirror neuron hypothesis and imitation impairments in autism spectrum disorders. *Neuroscience & Biobehavioral Reviews, 35*, 894–902. https://doi.org/10.1016/j.neubiorev.2010.10.007

Kehmeier, S., Schloegl, C., Scheiber, I. B. R., & Weiß, B. M. (2011). Early development of gaze following into distant space in juvenile Greylag geese (Anser anser). *Animal Cognition, 14*, 477–485. https://doi.org/10.1007/s10071-011-0381-x

Keven, N., & Akins, K. A. (2017). Neonatal imitation in context: Sensorimotor development in the perinatal period. *Behavioral and Brain Sciences, 40*, e381. https://doi.org/10.1017/S0140525X16000911

Keysar, B. (1994). The illusory transparency of intention: Linguistic perspective taking in text. *Cognitive Psychology, 26*, 165–208. https://doi.org/10.1006/cogp.1994.1006

Király, I., Jovanovic, B., Prinz, W., Aschersleben, G., & Gergely, G. (2003). The early origins of goal attribution in infancy. *Consciousness and Cognition: An International Journal, Self and Action, 12*, 752–769. https://doi.org/10.1016/S1053-8100(03)00084-9

Klein, K. J. K., & Hodges, S. D. (2001). Gender differences, motivation, and empathic accuracy: When it pays to understand. *Personality and Social Psychology Bulletin, 27*, 720–730. https://doi.org/10.1177/0146167201276007

Klinnert, M. D., Campos, J. J., Sorce, J. F., Emde, R. N., & Svejda, M. (1983). Emotions as behavior regulators: Social referencing in infancy. In R. Plutchik & H. Kellerman (Eds.), *Emotions in early development* (pp. 57–86). New York, NY: Academic Press. https://doi.org/10.1016/B978-0-12-558702-0.50009-1

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development, 76*, 1261–1277. https://doi.org/10.1111/j.1467-8624.2005.00849.x

Korman, J., & Malle, B. F. (2016). Grasping for traits or reasons? How people grapple with puzzling social behaviors. *Personality and Social Psychology Bulletin, 42*, 1451–1465. https://doi.org/10.1177/0146167216663704

Krauss, R. M., & Fussell, S. R. (1991). Perspective-taking in communication: Representations of others' knowledge in reference. *Social Cognition, 9*, 2–24. https://doi.org/10.1521/soco.1991.9.1.2

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science, 354*, 110. https://doi.org/10.1126/science.aaf8110

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Psychological Science, 14*, 402–408. https://doi.org/10.1111/1467-9280.01454

Lalonde, C. E., & Chandler, M. J. (2002). Children's understanding of interpretation. *New Ideas in Psychology, Folk Epistemology, 20*, 163–198. https://doi.org/10.1016/S0732-118X(02)00007-7

Latané, B., & Darley, J. M. (1968). Group inhibition of bystander intervention in emergencies. *Journal of Personality and Social Psychology, 10*, 215–221. https://doi.org/10.1037/h0026570

Lee, K., Eskritt, M., Symons, L. A., & Muir, D. (1998). Children's use of triadic eye gaze information for "mind reading". *Developmental Psychology, 34*, 525–539. https://doi.org/10.1037/0012-1649.34.3.525

Lee, R. B. (1972). Population growth and the beginnings of sedentary life among the !Kung bushmen. In B. Spooner (Ed.), *Population growth: Anthropological implications* (pp. 329–342). Cambridge, MA: MIT Press.

Leighton, J., Bird, G., Orsini, C., & Heyes, C. (2010). Social attitudes modulate automatic imitation. *Journal of Experimental Social Psychology, 46*, 905–910. https://doi.org/10.1016/j.jesp.2010.07.001

Lewis, M. (1997). The self in self-conscious emotions. In J. G. Snodgrass & R. L. Thompson (Eds.), *The self across psychology: Self-recognition, self-awareness, and the self-concept* (pp. 119–142). New York, NY: New York Academy of Sciences.

Lin, L. C., Qu, Y., & Telzer, E. H. (2018). Intergroup social influence on emotion processing in the brain. *Proceedings of the National Academy of Sciences, 115*, 10630–10635. https://doi.org/10.1073/pnas.1802111115

Lipps, T. (1907). Das Wissen von fremden Ichen. In T. Lipps (Ed.), *Psychologische Untersuchungen* (Vol. 1, pp. 641–693). Engelmann: Leipzig.

Louca-Papaleontiou, E., Melhuish, E., & Philaretou, A. (2012). Introspective abilities of preschool children. *Asian Transactions on Basic and Applied Sciences, 2*, 14–30.

Lundmark, M. (2017). *How to breathe life into inanimate objects*. Unpublished bachelor's thesis, Luleå University of Technology, Luleå, Sweden.

Luo, Y., & Choi, Y. (2013). Infants attribute mental states to nonhuman agents. In M. D. Rutherford & V. A. Kuhlmeier (Eds.), *Social perception: Detection and interpretation of animacy, agency, and intention* (pp. 259–281). Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262019279.003.0011

Lyons, D. E., Damrosch, D. H., Lin, J. K., Macris, D. M., & Keil, F. C. (2011). The scope and limits of overimitation in the transmission of artefact culture. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 366*, 1158–1167. https://doi.org/10.1098/rstb.2010.0335

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences, 104*, 19751–19756. https://doi.org/10.1073/pnas.0704452104

Malle, B. F. (2002). Verbs of interpersonal causality and the folk theory of mind and behavior. In M. Shibatani (Ed.), *The grammar of causation and interpersonal manipulation* (pp. 57–83). Amsterdam: Benjamins.

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press.

Malle, B. F. (2005). Folk theory of mind: Conceptual foundations of human social cognition. In R. R. Hassin, J. S. Uleman, & J. A. Bargh (Eds.), *The new unconscious* (pp. 225–255). New York, NY: Oxford University Press.

Malle, B. F. (2008). The fundamental tools, and possibly universals, of social cognition. In R. M. Sorrentino & S. Yamaguchi (Eds.), *Handbook of motivation and cognition across cultures* (pp. 267–296). New York, NY: Elsevier/Academic Press.

Malle, B. F. (2011a). Attribution theories: How people make sense of behavior. In D. Chadee (Ed.), *Theories in social psychology*. Wiley-Blackwell: Malden, MA.

Malle, B. F. (2011b). Time to give up the dogmas of attribution: A new theory of behavior explanation. In M. P. Zanna & J. M. Olson (Eds.), *Advances of experimental social psychology* (Vol. 44, pp. 297–352). San Diego, CA: Academic Press.

Malle, B. F. (2015). Social robots and the tree of social cognition. In Y. Nagai & S. Lohan (Eds.), *Proceedings of the Workshop "Cognition: A Bridge Between Robotics and Interaction" at HRI'15, Portland, Oregon* (pp. 13–14) Retrieved from http://www.macs.hw.ac.uk/~kl360/HRI2015W/proceedings.html

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*, 147–186. https://doi.org/10.1080/1047840X.2014.877340

Malle, B. F., & Holbrook, J. (2012). Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. *Journal of Personality and Social Psychology, 102*, 661–684. https://doi.org/10.1037/a0026790

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*, 101–121. https://doi.org/10.1006/jesp.1996.1314

Malle, B. F., & Knobe, J. (2001). The distinction between desire and intention: A folk-conceptual analysis. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 45–67). Cambridge, MA: MIT Press.

Malle, B. F., Knobe, J., & Nelson, S. E. (2007). Actor-observer asymmetries in explanations of behavior: New answers to an old question. *Journal of Personality and Social Psychology, 93*, 491–514. https://doi.org/10.1037/0022-3514.93.4.491

Margoni, F., & Surian, L. (2018). Infants' evaluation of prosocial and antisocial agents: A meta-analysis. *Developmental Psychology, 54*, 1445. https://doi.org/10.1037/dev0000538

Marsh, L. E., Ropar, D., & Hamilton, A. F. C. (2019). Are you watching me? The role of audience and object novelty in overimitation. *Journal of Experimental Child Psychology, 180*, 123. https://doi.org/10.1016/j.jecp.2018.12.010

Marshall, J. (1968). *Intention in law and society*. New York, NY: Funk & Wagnalls.

Measelle, J. R., John, O. P., Ablow, J. C., Cowan, P. A., & Cowan, C. P. (2005). Can children provide coherent, stable, and valid self-reports on the big five dimensions? A longitudinal study from ages 5 to 7. *Journal of Personality and Social Psychology, 89*, 90–106. https://doi.org/10.1037/0022-3514.89.1.90

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology, 31*, 838–850. https://doi.org/10.1037/0012-1649.31.5.838

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198*, 75–78. https://doi.org/10.1126/science.897687

Meltzoff, A. N., & Moore, M. K. (1997). Explaining facial imitation: A theoretical model. *Early Development & Parenting, Perceptual Development, 6*, 179–192. https://doi.org/10.1002/(SICI)1099-0917(199709/12)6:3/4<179::AID-EDP157>3.0.CO;2-R

Milgram, S., Bickman, L., & Berkowitz, L. (1969). Note on the drawing power of crowds of different size. *Journal of Personality and Social Psychology, 13*, 79–82. https://doi.org/10.1037/h0028070

Milgrom, P. R., North, D. C., & Weingast, B. R. (1990). The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs. *Economics and Politics, 2*, 1–23. https://doi.org/10.1111/j.1468-0343.1990.tb00020.x

Mitchell, J. P. (2006). Mentalizing and Marr: An information processing approach to the study of social cognition. *Brain Research, 1079*, 66–75. https://doi.org/10.1016/j.brainres.2005.12.113

Modelski, G. (2003). *World cities: -3000 to 2000*. Washington, DC: FAROS 2000.

Monroe, A. E., & Malle, B. F. (2017). Two paths to blame: Intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology: General, 146*, 123–133. https://doi.org/10.1037/xge0000234

Monroe, A. E., & Malle, B. F. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology, 116*, 215–236. https://doi.org/10.1037/pspa0000137

Moore, C., & Povinelli, D. J. (2007). Differences in how 12- and 24-month-olds interpret the gaze of adults. *Infancy, 11*, 215–231. https://doi.org/10.1111/j.1532-7078.2007.tb00224.x

Moore, D. (2011). Understanding of human motion, form and levels of meaning: Evidence from the perception of human point-light displays by infants and people with autism. In V. Slaughter & C. A. Brownell (Eds.), *Early development of body representations* (pp. 122–145). Cambridge: Cambridge University Press.

Moses, L. J., Baldwin, D. A., Rosicky, J. G., & Tidball, G. (2001). Evidence for referential understanding in the emotions domain at twelve and eighteen months. *Child Development, 72*, 718–735. https://doi.org/10.1111/1467-8624.00311

Myowa-Yamakoshi, M., Scola, C., & Hirata, S. (2012). Humans and chimpanzees attend differently to goal-directed actions. *Nature Communications, 3*, 1–7. https://doi.org/10.1038/ncomms1695

Nadel, J. (2002). Imitation and imitation recognition: Functional use in preverbal infants and nonverbal children with autism. In A. N. Meltzoff & W. Prinz (Eds.), *The imitative mind: Development, evolution, and brain bases* (pp. 42–62). New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511489969.003

Neumann, R., & Strack, F. (2000). "Mood contagion": The automatic transfer of mood between persons. *Journal of Personality and Social Psychology, 79*, 211–223. https://doi.org/10.1037/0022-3514.79.2.211

Nielsen, M. (2018). The social glue of cumulative culture and ritual behavior. *Child Development Perspectives, 12*, 264. https://doi.org/10.1111/cdep.12297

Nielsen, M., Mushin, I., Tomaselli, K., & Whiten, A. (2014). Where culture takes hold: "Overimitation" and its flexible deployment in Western, Aboriginal, and Bushmen children. *Child Development, 85*, 2169–2184.

Nielsen, M., & Tomaselli, K. (2010). Overimitation in Kalahari bushman children and the origins of human cultural cognition. *Psychological Science, 21*, 729–736. https://doi.org/10.1177/0956797610368808

O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Development, 67*, 659–677. https://doi.org/10.2307/1131839

Ochsner, K. (2013). The role of control in emotion, emotion regulation, and empathy. In D. Hermans, B. Rimé, & B. Mesquita (Eds.), *Changing emotions* (pp. 157–165). New York, NY: Psychology Press.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*, 255–258. https://doi.org/10.1126/science.1107621

Oostenbroek, J., Suddendorf, T., Nielsen, M., Redshaw, J., Kennedy-Costantini, S., Davis, J., et al. (2016). Comprehensive longitudinal study challenges the existence of neonatal imitation in humans. *Current Biology, 26*, 1334–1338. https://doi.org/10.1016/j.cub.2016.03.047

Osterhaus, C., Koerber, S., & Sodian, B. (2016). Scaling of advanced theory-of-mind tasks. *Child Development, 87*, 1971–1991. https://doi.org/10.1111/cdev.12566

Palagi, E., Norscia, I., Pressi, S., & Cordoni, G. (2019). Facial mimicry and play: A comparative study in chimpanzees and gorillas. *Emotion, 19*, 665. https://doi.org/10.1037/emo0000476

Peregrine, P. N., Ember, C. R., & Ember, M. (2007). Modeling state origins using cross-cultural data. *Cross-Cultural Research, 41*, 75–86. https://doi.org/10.1177/1069397106295445

Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science (New York, N.Y.), 308*, 214–216. https://doi.org/10.1126/science.1111656

Persson, T., Sauciuc, G.-A., & Madsen, E. A. (2018). Spontaneous cross-species imitation in interactions between chimpanzees and zoo visitors. *Primates, 59*, 19–29. https://doi.org/10.1007/s10329-017-0624-9

Poulin-Dubois, D., Brooker, I., & Chow, V. (2009). The developmental origins of naïve psychology in infancy. *Advances in Child Development and Behavior, 37*, 55–104.

Povinelli, D. J., & Eddy, T. J. (2000). *What young chimpanzees know about seeing*. New York, NY: Wiley-Blackwell.

Povinelli, D. J., Landau, K. R., & Perilloux, H. K. (1996). Self-recognition in young children using delayed versus live feedback: Evidence of a developmental asynchrony. *Child Development, 67*, 1540–1554. https://doi.org/10.2307/1131717

Povinelli, D. J., & Preuss, T. M. (1995). Theory of mind: Evolutionary history of a cognitive specialization. *Trends in Neurosciences, 18*, 418–424. https://doi.org/10.1016/0166-2236(95)93939-U

Povinelli, D. J., Rulf, A. B., Landau, K. R., & Bierschwale, D. T. (1993). Self-recognition in chimpanzees (Pan troglodytes): Distribution, ontogeny, and patterns of emergence. *Journal of Comparative Psychology, 107*, 347–372. https://doi.org/10.1037/0735-7036.107.4.347

Povinelli, D. J., & Vonk, J. (2003). Chimpanzee minds: Suspiciously human? *Trends in Cognitive Sciences, 7*, 157–160. https://doi.org/10.1016/S1364-6613(03)00053-6

Powell, P. A. (2018). Individual differences in emotion regulation moderate the associations between empathy and affective distress. *Motivation and Emotion, 42*, 602. https://doi.org/10.1007/s11031-018-9684-4

Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition, 36*, 1–16. https://doi.org/10.1016/0010-0277(90)90051-K

Prochazkova, E., & Kret, M. E. (2017). Connecting minds and sharing emotions through mimicry: A neurocognitive model of emotional contagion. *Neuroscience & Biobehavioral Reviews, 80*, 99–114. https://doi.org/10.1016/j.neubiorev.2017.05.013

Qureshi, A. W., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: Evidence from a dual-task study of adults. *Cognition, 117*, 230–236. https://doi.org/10.1016/j.cognition.2010.08.003

Ramon, M., Caharel, S., & Rossion, B. (2011). The speed of recognition of personally familiar faces. *Perception, 40*, 437–449. https://doi.org/10.1068/p6794

Rauchbauer, B., Majdandžić, J., Stieger, S., & Lamm, C. (2016). The modulation of mimicry by ethnic group-membership and emotional expressions. *PLoS One, 11*, e0161064. https://doi.org/10.1371/journal.pone.0161064

Ray, E., & Heyes, C. (2011). Imitation in infancy: The wealth of the stimulus. *Developmental Science, 14*, 92–105. https://doi.org/10.1111/j.1467-7687.2010.00961.x

Redman, C. L. (1978). *The rise of civilization: From early farmers to urban society in the ancient Near East*. San Francisco, CA: W. H. Freeman.

Renfrew, C. (2007). *Prehistory: The making of the human mind*. London: Weidenfeld & Nicolson.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology, 33*, 12–21. https://doi.org/10.1037/0012-1649.33.1.12

Ricaut, F.-X., Cox, M. P., Lacan, M., Keyser, C., Duranthon, F., Ludes, B., et al. (2012). A time series of prehistoric mitochondrial DNA reveals Western European genetic diversity was largely established by the Bronze Age. *Advances in Anthropology, 2*, 14–23. https://doi.org/10.4236/aa.2012.21002

Ross, L., & Nisbett, R. E. (1991). *The person and the situation*. New York, NY: McGraw-Hill.

Roth-Hanania, R., Davidov, M., & Zahn-Waxler, C. (2011). Empathy development from 8 to 16 months: Early signs of concern for others. *Infant Behavior & Development, 34*, 447–458. https://doi.org/10.1016/j.infbeh.2011.04.007

Ruffman, T., Aitken, J., Wilson, A., Puri, A., & Taumoepeau, M. (2018). A re-examination of the broccoli task: Implications for children's understanding of subjective desire. *Cognitive Development, 46*, 79. https://doi.org/10.1016/j.cogdev.2017.08.001

Ruffman, T., Lorimer, B., & Scarf, D. (2017). Do infants really experience emotional contagion? *Child Development Perspectives, 11*, 270–274. https://doi.org/10.1111/cdep.12244

Ruffman, T., & Taumoepeau, M. (2014). When and how does a theory of mind arise? In O. N. Saracho (Ed.), *Contemporary perspectives on research in theory of mind in early childhood education* (pp. 45–68). Charlotte, NC: IAP Information Age Publishing.

Ruffman, T., Then, R., Cheng, C., & Imuta, K. (2019). Lifespan differences in emotional contagion while watching emotion-eliciting videos. *PLoS One, 14*, e0209253. https://doi.org/10.1371/journal.pone.0209253

Sabbagh, M. A., & Paulus, M. (2018). Replication studies of implicit false belief with infants and toddlers. *Cognitive Development, Understanding Theory of Mind in Infancy and Toddlerhood, 46*, 1–3. https://doi.org/10.1016/j.cogdev.2018.07.003

Sato, W., & Yoshikawa, S. (2007). Spontaneous facial mimicry in response to dynamic facial expressions. *Cognition, 104*, 1–18. https://doi.org/10.1016/j.cognition.2006.05.001

Saylor, M. M., Baldwin, D. A., Baird, J. A., & LaBounty, J. (2007). Infants' on-line segmentation of dynamic human action. *Journal of Cognition and Development, 8*, 113–128. https://doi.org/10.1207/s15327647jcd0801_6

Scambler, D. J., Hepburn, S., Rutherford, M. D., Wehner, E. A., & Rogers, S. J. (2007). Emotional responsivity in children with autism, children with other developmental disabilities, and children with typical development. *Journal of Autism and Developmental Disorders, 37*, 553–563. https://doi.org/10.1007/s10803-006-0186-y

Schult, C. A. (2002). Children's understanding of the distinction between intentions and desires. *Child Development, 73*, 1727–1747.

Shaver, K. G. (1975). *An introduction to attribution processes*. Cambridge, MA: Winthrop.

Shostak, M. (2009). *Nisa: The life and words of a !Kung woman*. Cambridge, MA: Harvard University Press.

Shultziner, D., Stevens, T., Stevens, M., Stewart, B. A., Hannagan, R. J., & Saltini-Semerari, G. (2010). The causes and scope of political egalitarianism during the Last Glacial: A multidisciplinary perspective. *Biology and Philosophy, 25*, 319–346. https://doi.org/10.1007/s10539-010-9196-4

Sirois, S., & Jackson, I. (2007). Social cognition in infancy: A critical review of research on higher order abilities. *European Journal of Developmental Psychology, 4*, 46–64. https://doi.org/10.1080/17405620601047053

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105*, 131–142. https://doi.org/10.1037//0033-2909.105.1.131

Slaughter, V., & McConnell, D. (2003). Emergence of joint attention: Relationships between gaze following, social referencing, imitation, and naming in infancy. *The Journal of Genetic Psychology, 164*, 54–71. https://doi.org/10.1080/00221320309597503

Sloman, S., & Fernbach, P. (2017). *The knowledge illusion: Why we never think alone*. New York, NY: Riverhead Books.

Sommerville, J. A., Woodward, A. L., & Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cognition, 96*, B1–B11. https://doi.org/10.1016/j.cognition.2004.07.004

Song, H., Onishi, K. H., Baillargeon, R., & Fisher, C. (2008). Can an agent's false belief be corrected by an appropriate communication? Psychological reasoning in 18-month-old infants. *Cognition, 109*, 295–315. https://doi.org/10.1016/j.cognition.2008.08.008

Sonnby-Borgström, M., Jönsson, P., & Svensson, O. (2003). Emotional empathy as related to mimicry reactions at different levels of information processing. *Journal of Nonverbal Behavior, 27*, 3–23. https://doi.org/10.1023/A:1023608506243

Sorce, J. F., Emde, R. N., Campos, J. J., & Klinnert, M. D. (1985). Maternal emotional signaling: Its effect on the visual cliff behavior of 1-year-olds. *Developmental Psychology, 21*, 195–200. https://doi.org/10.1037/0012-1649.21.1.195

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18*, 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Stel, M., Van Baaren, R. B., & Vonk, R. (2008). Effects of mimicking: Acting prosocially by being emotionally moved. *European Journal of Social Psychology, 38*, 965–976. https://doi.org/10.1002/ejsp.472

Stel, M., & Vonk, R. (2010). Mimicry in social interaction: Benefits for mimickers, mimickees, and their interaction. *British Journal of Psychology, 101*, 311–323. https://doi.org/10.1348/000712609X465424

Sterck, E. H. M., & Begeer, S. (2010). Theory of mind: Specialized capacity or emergent property? *European Journal of Developmental Psychology, 7*, 1–16. https://doi.org/10.1080/17405620903526242

Stiles, R. J. (1985). *Mimicry as a mechanism of arousal of empathy in children*. Unpublished master thesis, University of Michigan, Ann Arbor, MI.

Subiaul, F., Renner, E., & Krajkowksi, E. (2016). The comparative study of imitation mechanisms in non-human primates. In S. S. Obhi & E. S. Cross (Eds.), *Shared representations: Sensorimotor foundations of social life. Cambridge social neuroscience* (pp. 109–135). New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9781107279353.007

Suddendorf, T., & Butler, D. L. (2013). The nature of visual self-recognition. *Trends in Cognitive Sciences, 17*, 121–127. https://doi.org/10.1016/j.tics.2013.01.004

Sutcliffe, A., Dunbar, R., Binder, J., & Arrow, H. (2012). Relationships and the social brain: Integrating psychological and evolutionary perspectives. *British Journal of Psychology, 103*, 149–168. https://doi.org/10.1111/j.2044-8295.2011.02061.x

Taylor, M., Esbensen, B. M., & Bennett, R. T. (1994). Children's understanding of knowledge acquisition: The tendency for children to report that they have always known what they have just learned. *Child Development, 65*, 1581–1604. https://doi.org/10.2307/1131282

Thomas, F., & Johnston, O. (1995). *The illusion of life: Disney animation*. New York, NY: Hyperion.

Tomasello, M. (1998). Social cognition and the evolution of culture. In J. Langer & M. Killen (Eds.), *Piaget, evolution, and development* (pp. 221–245). Mahwah, NJ: Lawrence Erlbaum Associates.

Tomasello, M., & Carpenter, M. (2005). The emergence of social cognition in three young chimpanzees. *Monographs of the Society for Research in Child Development, 70*, 1–122. https://doi.org/10.1111/j.1540-5834.2005.00330.x

Tomasello, M., & Rakoczy, H. (2003). What makes human cognition unique? From individual to shared to collective intentionality. *Mind & Language, 18*, 121–147. https://doi.org/10.1111/1468-0017.00217

Tremlin, T. (2006). *Minds and gods: The cognitive foundations of religion*. New York, NY: Oxford University Press.

Van Boven, L., & Loewenstein, G. (2003). Social projection of transient drive states. *Personality and Social Psychology Bulletin, 29*, 1159–1168. https://doi.org/10.1177/0146167203254597

Van der Borght, M., & Jaswal, V. K. (2009). Who knows best? Preschoolers sometimes prefer child informants over adult informants. *Infant and Child Development, 18*, 61–71. https://doi.org/10.1002/icd.591

Van Overwalle, F., Van Duynslaeger, M., Coomans, D., & Timmermans, B. (2012). Spontaneous goal inferences are often inferred faster than spontaneous trait inferences. *Journal of Experimental Social Psychology, 48*, 13–18. https://doi.org/10.1016/j.jesp.2011.06.016

Vincini, S., Jhang, Y., Buder, E. H., & Gallagher, S. (2017). Neonatal imitation: Theory, experimental design, and significance for the field of social cognition. *Frontiers in Psychology, 8*, 1323. https://doi.org/10.3389/fpsyg.2017.01323

Voiklis, J., & Malle, B. F. (2017). Moral cognition and its basis in social cognition and social regulation. In K. Gray & J. Graham (Eds.), *Atlas of moral psychology*. New York, NY: Guilford Press.

Walden, T. A., & Ogan, T. A. (1988). The development of social referencing. *Child Development, 59*, 1230–1240. https://doi.org/10.2307/1130486

Walle, E. A., Reschke, P. J., & Knothe, J. M. (2017). Social referencing: Defining and delineating a basic process of emotion. *Emotion Review, 9*, 245–252. https://doi.org/10.1177/1754073916669594

Wang, Y., & Hamilton, A. F. C. (2012). Social top-down response modulation (STORM): A model of the control of mimicry in social interaction. *Frontiers in Human Neuroscience, 6*, 153. https://doi.org/10.3389/fnhum.2012.00153

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*, 655–684. https://doi.org/10.1111/1467-8624.00304

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75*, 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

Wellman, H. M., & Phillips, A. T. (2001). Developing intentional understandings. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 125–148). Cambridge, MA: The MIT Press.

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition, 35*, 245–275.

Wen, N. J., Herrmann, P. A., & Legare, C. H. (2016). Ritual increases children's affiliation with in-group members. *Evolution and Human Behavior, 37*, 54–60. https://doi.org/10.1016/j.evolhumbehav.2015.08.002

Williams, E. (1985). Estimation of prehistoric populations of archaeological sites in southwestern Victoria: Some problems. *Archaeology in Oceania, 20*, 73–80.

Wilson, P. J. (1988). *The domestication of the human species*. New Haven, CT: Yale University Press.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69*, 1–34. https://doi.org/10.1016/S0010-0277(98)00058-4

Woodward, A. L. (2003). Infants' developing understanding of the link between looker and object. *Developmental Science, 6*, 297–311. https://doi.org/10.1111/1467-7687.00286

Yu, L., Hattori, Y., Yamamoto, S., & Tomonaga, M. (2018). Understanding empathy from interactional synchrony in humans and non-human primates. In L. D. Di Paolo, F. Di Vincenzo, & F. De Petrillo (Eds.), *Evolution of primate social cognition* (pp. 47–58). Cham: Springer. https://doi.org/10.1007/978-3-319-93776-2_4

Zeder, M. A. (2011). Religion and the revolution: The legacy of Jacques Cauvin. *Paléorient, 37*, 39–60. https://doi.org/10.3406/paleo.2011.5437

The Cognitive Basis of Mindreading

Ian Apperly

Why did Anna Karenin throw herself under a train? A satisfactory answer to this question will surely refer to Anna's mental states—her thoughts and feelings, desires, and intentions. Most readers of Tolstoy's novel would consider such mindreading essential to understanding the story. They might also find that an important part of Tolstoy's craft is the generation of tension between Anna's perspective and emotional state and those of other characters, and their own perspective as a reader. It is deeply revealing about the nature of mindreading that we find it quite natural to think about the mental states of a fictional character, from a different place and time, in an unusual set of personal circumstances, and this exposes important limitations of common claims and assumptions about mindreading.

Neuroscientific approaches have much to teach us about the nature of mindreading but, as in other areas of cognitive neuroscience, they are at their most powerful when combined with clear hypotheses about the cognitive processes involved. I begin by considering the limitations of some prominent theoretical ideas about mindreading. I will go on to describe a cognitive account that, I think, provides a better foundation for a cognitive neuroscience of mindreading. I will highlight examples of what neuroscientific approaches have already told us about the cognitive basis of mindreading, before considering some exciting future prospects.

Mindreading Mantras

Mindreading has been extensively theorized by psychologists, linguists, and philosophers. This offers a rich inheritance to empirical investigators. However, bold conjectures about how mindreading might work have sometimes become received

I. Apperly (✉)
School of Psychology, University of Birmingham, Birmingham, UK
e-mail: i.a.apperly@bham.ac.uk

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_18

wisdom about how it does work or must work, which can cloud our thinking about what mindreading is and how to study it. To persuade you that it's worth engaging seriously with questions about the cognitive basis of mindreading, let me challenge some oft-repeated claims.

Mindreading is not just "decoding" of mental states from behaviour. It is commonly assumed that mental states can be decoded from behaviour, in much the same way as words can be decoded from text (e.g., Heyes, 2018). Of course it is true that being able to interpret a facial expression as evidence of an emotion, or search behaviour as evidence of a belief about an object's location and a desire to find it, are important components of mindreading. Equally, however, such decoding is not the essence of mindreading. It is clear from the example of Anna Karenin that we may mindread without direct perceptual access to behaviour. Moreover, many of the mental states we might ascribe to Anna—such as her anxiety about her social position—follow from facts about her background, about other characters, or the context, none of which we have observed. Mindreading real people is no different. Moreover, Tolstoy sometimes simply tells us what Anna is thinking; just real people sometimes report on their own mental states, and those of others. Therefore, while observed behaviour is surely one important input for mindreading, it is not necessary and, other than in the simplest cases, it is not usually sufficient.

People do not have a "theory" of mind It is commonly claimed that our mindreading abilities consist in theory, involving concepts—"belief", "desire", "intention", etc.—and principles for how they combine (e.g. Davies & Stone, 1995; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). Just as someone who knows the words and grammar of a language is equipped to parse sentences of that language, so someone with a "theory of mind" would be able to use mental states for explanations or predictions of behaviour. However, unlike linguistic grammars, 30 years of research on mindreading has not codified the supposed principles by which mental states interact for realistic scenarios (Stuhlmüller & Goodman, 2014). There is no extant theory that can parse Anna's circumstances into a reliable set of thoughts and feelings. Instead there are good grounds for supposing that the complexity of the interactions among mental states and between mental states and behaviour is uncodifiable (e.g. Davidson, 1990). Note that this should not be taken as support for "simulation" accounts of mindreading, which do not offer easy solutions to this problem (e.g. Apperly, 2008).

Mindreading does not make unique reasoning demands. An influential early account suggested that mindreading poses unique logical problems, which require a unique representational solution (e.g. Leslie, 1987). This strong hypothesis is difficult to sustain since similar logical problems arise when we need to set aside our own current situation to reason about different times, places, or counterfactuals (Barwise & Perry, 1983; Fauconnier, 1985). Moreover, there are empirical associations between mindreading tasks and non-mindreading tasks that are matched in their logical and structural requirements (Perner & Leekam, 2008). A reasonable conclusion from such work is that mindreading poses some exacting representational challenges, but not unique ones (Apperly, 2010).

Mindreading is not simply automatic What people mean when they claim that mindreading is automatic seems to range from the intuition that mindreading is natural and effortless to a firm commitment to mindreading being a quasi-perceptual Fodor-module (e.g. Leslie, 2005; Stone, Baron-Cohen, & Knight, 1998). Either way, direct investigations have provided evidence that mindreading meets important criteria for automaticity in some circumstances (e.g. Kovács, Téglás, & Endress, 2010; Qureshi, Apperly, & Samson, 2010; Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010; van der Wel, Sebanz, & Knoblich, 2014), but shows clear non-automaticity in others (e.g. Apperly, Riggs, Simpson, Chiavarino, & Samson, 2006). While some of these results remain controversial (e.g. Heyes, 2014; Phillips et al., 2015), it has been suggested that they can be reconciled in a "two systems" account, whereby humans have the capacity to make a minimal set of mindreading inferences automatically, and a second ability that is more effortful but flexible enough to cope with the complexity of full-blown mindreading (e.g. Apperly & Butterfill, 2009; Low, Apperly, Butterfill, & Rakoczy, 2016). The latter would be key to mindreading Anna Karenin, where, on analogy with other inferences made during comprehension, mindreading would be spontaneous (i.e. uninstructed) but conditional on having the requisite processing resources and the motivation to use them (Apperly, 2010).

In summary, mindreading involves much more than "decoding" mental states from behaviour, not least because there is nothing like an exhaustive code. Mindreading makes similar representational demands to structurally similar problems that have nothing to do with mindreading. Only some kinds of mindreading judgement show signs of automaticity; others—such as the problem of figuring out why Anna Karenina threw herself under a train—are clearly effortful and contingent on resources and motivation. In functional terms, such mindreading requires complex, flexible processing over our full database of knowledge about the world, and so fits Fodor's criteria for "central" rather than "modular" processes (Fodor, 1983). From this perspective, it should be no surprise to discover that mindreading involves a rich set of processes for representation, reasoning and control, supported by a network of brain regions. However, this also demands some kind of functional model to organize existing findings and guide new research. Below I summarize such a model. A fuller justification in terms of empirical findings can be found in Apperly (2010).

A Cognitive Model of Mindreading

The great majority of mindreading tasks present a situation involving an agent with mental states that differ from participants'. The agent's mental states must be inferred and either reported, or else used to predict their subsequent behaviour. In doing so these tasks combine and confound most of the functional processes that contribute to mindreading. If we only used this approach, it would be like trying to understand the cognitive and neural basis of language by only ever presenting participants with tasks that combine every level of phonological, syntactic and semantic processing in a full cycle of comprehension and production. To break out of this problem, we need theories of the functional components of mindreading and tasks that allow putative functional components to be distinguished.

I find box and arrow models extremely useful for organizing ideas about the cognitive basis of mindreading. In the following, I focus on mindreading what someone thinks or knows. The level of description is "computational" in Marr's sense (Marr, 1982), so model components say something about what the system is doing, with no commitment to the algorithmic or neural implementation of those functions. However, such a model of these functions is, I think, essential for explaining or predicting the demands made by different mindreading tasks and how these affect the recruitment of neural systems during mindreading.

The horizontal dimension of Fig. 1 distinguishes the need to infer what someone else is thinking from the need to store this information, and from the use of this information to predict or explain what someone is doing or saying. The vertical dimension distinguishes between "system 1" (below) and "system 2" (above) processes (e.g. Evans & Stanovich, 2013). System 2 processes enable highly flexible mindreading, so in principle I could ascribe to you or to Anna Karenin any thought that I could entertain for myself. However, this flexibility comes at the expense of System 2 thinking making higher demands than System 1 on scarce resources for memory and cognitive control (e.g. Low et al., 2016). System 1 trades reduced flexibility for increased efficiency. Increased efficiency is evidenced in the apparent automaticity with which some mindreading processes occur, and limited dependence on cognitive control processes (e.g. Kovács et al., 2010; Qureshi et al., 2010;

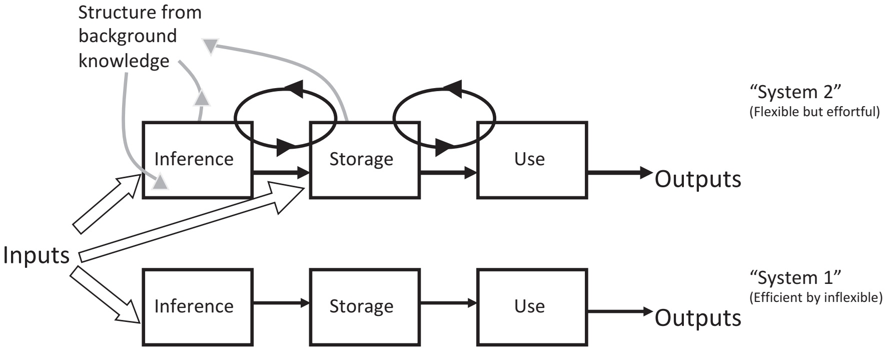


Fig. 1 A "two systems" model of mindreading (simplified from Apperly, 2010). The model distinguishes processes involved in inference, storage and use of information about others' mental states. "System 2" makes flexible, context-sensitive mindreading inferences by drawing richly upon background knowledge, in processes represented by the grey arrows. Oval arrows indicate that System 2 mindreading will often involve repeated cycles of reasoning. System 1 processes manage to be more cognitively efficient by limiting their interaction with background information and limiting their processing over inputs. For clarity, only one System 1 process is depicted, but there are likely to be multiple processes, for example to enable mindreading of belief-like states, goals and emotions

Samson et al., 2010; van der Wel et al., 2014), while reduced flexibility is evidenced by the appearance of automaticity only for relatively simple problems (such as inferring what someone sees, Samson et al., 2010) and not for more complex problems (such as inferring precisely how they see it from their perspective; Surtees, Samson, & Apperly, 2016). The greater number and complexity of arrows for System 2 reflects the greater flexibility of information flow compared with System 1. The main focus of the present chapter will be on System 2 processes.

Implicit in Fig. 1 is the fact that mindreading requires the representation of someone else and their mental states as distinct from one's self and one's own. Maintaining this distinction is essential, but it is also challenging because the perspective of the other and of one's self are not independent records of facts, but are related to each other and to "reality". This closely related information gives rise to interference between self and other, such that if I represent our differing beliefs about something (even something as mundane as the location of a hidden object), I am slower and more error-prone when judging what you think ("egocentric interference", Royzman, Cassidy, & Baron, 2003), and when judging what I think myself ("altercentric interference", Samson et al., 2010). A successful mindreader must not only maintain a distinction between the perspectives of self and other, but also manage the interference that results: mindreading requires inference, representation and control.

Figure 1 helps systematize a set of important questions about mindreading. For example, are any or all of these processes specialized for mindreading; do the cognitive control requirements arise at all stages of processing; is the network of brain regions implicated in mindreading equally involved in inference, storage and use of mindreading information? In the next section I will tackle some of these questions, and show how a cognitive model helps us understand what light cognitive neuroscience has already shone on our understanding of mindreading.

Specialization for Mindreading

While mindreading does not appear to make unique reasoning demands, a related hypothesis is that the cognitive and neural systems for mindreading are domain-specific. The latter does not entail the former, because reasons other than unique reasoning demands could lead mindreading to show domain specificity. For example, if there is neural specialization for other social processes (e.g. Adolphs, 2009; Frith, 2007) neural activity during mindreading may show domain specificity for at least three reasons: (1) because one or more of those other social processes are intrinsic mindreading, (2) because those social processes have distinctive neural connectivity with neural systems involved in mindreading, (3) because mature mindreading develops on the foundation of other social processes that are themselves domain-specific, and so inherits domain specificity without this being functionally necessary.

Domain specificity for mindreading has been tested most extensively in a series of studies by Saxe and colleagues (Koster-Hale & Saxe, 2013; Saxe & Kanwisher, 2003). This widely adopted approach starts by contrasting neural activation while participants reason about false beliefs with activation during structurally and logically similar reasoning about false photographs and false signs. Brain activity surviving this contrast is then tested for its selectivity for a range of other judgements about people's mental states, personality, physical appearance and other characteristics. While the contrast between false beliefs and false photos typically reveals activity in mPFC, bilateral TPJ and temporal poles, over an impressive range of studies it is right TPJ that shows the highest selectivity for reasoning about mental states (Koster-Hale & Saxe, 2013).

These results illustrate the value of cognitive neuroimaging for understanding mindreading because they provide stronger evidence than behavioural studies that mindreading involves domain-specific processes. However, there are also important caveats. First, demonstrating domain specificity is just one step towards understanding underlying mechanisms, and for now it remains unclear what function rTPJ is performing or what feature of mindreading leads to evidence of domain specificity (see Future Prospect, below). It is not clear whether domain-specific processes are involved in inference, storage or use of mindreading information, or all three (Fig. 1). Second, as described above, it is clear that mindreading depends upon many processes, which will not all be domain-specific. It's therefore important that questions about domain specificity are complemented by questions about the broader functional basis of mindreading. Third, the best methods for testing the domain specificity of mindreading are unsuitable for understanding these broader components of mindreading because such processes are subtracted out of the comparison between strictly matched mindreading and non-mindreading tasks. The most obvious examples of this are processes involved in the control of mindreading.

Control Processes During Mindreading

Control of egocentrism A vivid illustration that domain-general processes contribute significantly to mindreading comes from the neuropsychological case study of patient WBA (Samson, Apperly, Kathirgamanathan, & Humphreys, 2005). This patient sustained a right frontal brain lesion, following a stroke, and his lesion affected lateral frontal brain regions most commonly implicated in cognitive control, notably including right inferior frontal gyrus. Consistent with this WBA showed notable impairment on standard neuropsychological assessments of executive function, including inhibitory control. Medial PFC—commonly implicated in mindreading—was left largely intact. Consistent with this, WBA appeared to be able to reason about other people's false beliefs, provided he was tested on an unusual task that minimized the salience of his own knowledge of the correct answer. However, on more standard false belief tasks and on a range of other tests of his ability to judge other people's perspectives he showed very high rates of "egocentric errors", where he responded according to his own perspective rather than the other person's. Anecdotal report from a family member indicated that this pronounced egocentrism was not limited to laboratory tasks.

Importantly, such egocentric errors are not simply the product of generic task difficulty. In a follow-up study WBA, and another patient with similar brain injury, showed egocentric errors when required to judge the differing desires of an opponent in a card game, but lower errors when judging the card they next needed themselves, despite variation in whether a matching or a mismatching card would be a winner. A second pair of patients with lesions to more medial prefrontal cortex showed the opposite pattern of errors (Samson, Houthuys, & Humphreys, 2015). This demonstrates a classical neuropsychological double dissociation between the control processes necessary for managing interference from self perspective when taking the other's perspective, versus those necessary for handling conflict arising from other aspects of game strategy.

Such evidence from studies of patients converges with evidence from fMRI, ERP and TMS in suggesting a selective role for lateral frontal regions—in particular inferior frontal gyrus—in controlling tendencies for both egocentric and altercentric error and bias during mindreading (e.g. McCleery, Surtees, Graham, Richards, & Apperly, 2011; van der Meer et al., 2011; Vogeley et al., 2001). For example, Hartwright, Apperly, and Hansen (2012) used a "belief-desire" task in which participants used a character's beliefs and desires to predict their search in one of two boxes. Participants were told which box contained some food, which box the character thought contained the food, and whether or not the character desired the food on that trial. When the character's belief was false there was conflict between his perspective and the participants', but not when his belief was true. In contrast the character's desire for the food was not systematically related to the participants' (he might like peas, whereas the participant does not), so conflict was equally likely to occur (or not occur) at each level of this factor. Consistent with previous behavioural studies (e.g. Apperly, Warren, Andrews, Grant, & Todd, 2011; German & Hehman, 2006), responses were slower and more error-prone whenever the character's belief was false and whenever his desire was negative. A natural interpretation of these results might be that the belief and desire effects were equivalent, perhaps because false belief and negative desire both required more inhibitory control (Friedman & Leslie, 2004). However, fMRI data suggested that these effects were not equivalent: whereas activity in bilateral TPJ and dorsomedial PFC was influenced by both belief and desire, activity in right IFG was influenced only by the factor of belief, and not by the factor of desire. Moreover, in a subsequent study, r-TMS to right IFG influenced performance on false versus true belief trials, and not negative versus positive desire trials (Hartwright, Hardwick, Apperly, & Hansen, 2016). These findings converge with the neuropsychological evidence in suggesting that IFG is involved specifically in resisting "egocentric" interference from self perspective when taking the perspective of someone else.

Self versus other The need to control interference from self perspective when taking the perspective of another presupposes that you have represented the other's perspective. In parallel with work on controlling egocentrism is a burgeoning literature on the cognitive and neural basis of distinguishing self from other (e.g. Cook, 2014). This work began with evidence that observing another's action creates a tendency for "automatic imitation" of the action by one's self, which must be controlled if a different action is necessary for the task (Brass, Bekkering, Wohlschläger, & Prinz, 2000). Whereas controlling interference from other kinds of over-learned association is typically linked with activity in lateral prefrontal brain regions (e.g. Wagner, Maril, Bjork, & Schacter, 2001), control of automatic imitation appears to depend on regions of mPFC and TPJ similar or identical to those commonly implicated in mindreading. A number of studies suggest that this link with mindreading is more than coincidental. For example, Santiesteban, White, et al. (2012) found that training inhibition of automatic imitation improved participants' use of mindreading in a communication task (the Director Task; Keysar, Lin & Barr, 2003), whereas training generic inhibition did not. Santiesteban, Banissy, Catmur, and Bird (2012) found that stimulation of rTPJ improved both inhibition of imitation and use of mindreading in a communication task. Such findings suggest that the same process of self-other control may be at work in both imitation inhibition and perspective-taking, with one hypothesis being that TPJ maintains the distinction between information related to self versus other (perhaps in line with its role in general control of attention), while mPFC prioritizes one or other set of information according to the task or the context (Santiesteban, Banissy, et al., 2012).

An assimilation On the face of it, these data appear contradictory to those presented in the previous section: "self/other control" and "control of egocentrism" sound a lot like two terms for the same phenomenon, yet the data suggest they depend on different functional and neural processes. I suggest, however, that if we think about mindreading in terms of component processes then there may be no contradiction. An example will help illustrate the point. McCleery et al. (2011) used a simple perspective-taking task in which participants viewed a schematic room with dots on the wall and an avatar standing in the middle. The avatar's position meant that the number of dots he saw was sometimes consistent with the participant's perspective and sometimes inconsistent. On some trials participants were told to judge how many dots they themselves saw when the picture appeared (self trials) while on other trials they judged how many the avatar saw (other trials). Participants are slower to judge both self and other perspectives whenever those perspectives are inconsistent (Samson et al., 2010), and a simultaneous executive task increases this effect to an equal degree for self and other judgements (Qureshi et al., 2010). We have interpreted this pattern to suggest that self and other perspectives are calculated on every trial in a relatively effortless manner ("inference" in Fig. 1), with the effortful step being a subsequent process of selecting either self or other perspective as the basis for a response ("use" in Fig. 1). McCleery et al. (2011) recorded ERPs during this task. They observed a component from electrodes over temporoparietal cortex approximately 450 ms after picture onset, which varied according to whether participants were making self or other judgements. They also observed a later and longer-lasting component from electrodes over right frontal cortex, which varied only according to whether self and other perspectives were consistent versus inconsistent. These effects were tentatively localized to left and right TPJ and right IFG, respectively.

I suggest that these results support generalizable conclusions that help make sense of a variety of findings about control processes during mindreading. Mindreading requires the establishment and maintenance of a distinction between self and other, which depends on TPJ. This may well be necessary at all processing steps: inferences, storage and use (Fig. 1). Having distinguished self and other we are then in a position to use either self or other perspective to make responses or to inform further processing. Whichever perspective we are trying to use, the other perspective will tend to compete, potentially activating the response relevant to the opposite perspective from the one intended. This latter interference may originate with representations of perspectives but in other respects resembles entirely generic interference effects and recruits generic processes associated with IFG. It occurs most clearly during the use of mindreading information (Fig. 1), and difficulty with resisting this interference leads to a large number of the egocentric phenomena reported in the literature.

Mindreading Inferences

While almost all mindreading tasks require participants to infer a target's mental states, Fig. 1 encourages us to distinguish such inferences from other mindreading processes. And just as Tolstoy can tell us what Anna is thinking, and real people can inform us of their thoughts and feelings, so we can create experimental tasks that remove the need to infer the mental states of others. Among other things, this allows us to ask whether any brain areas involved in mindreading are distinctively involved in such mindreading inferences. The belief-desire task described earlier (Apperly et al., 2011; Hartwright et al., 2012) opens this possibility, because participants are simply told the character's belief and desire. In terms of Fig. 1, participants skip the initial inference step, but must *store* the mental states they are told and *use* them to reason about the character's behaviour. Hartwright et al. also employed the false belief/false photograph "localizer" task developed by Saxe and colleagues, which clearly does involve mindreading inferences. In the belief-desire task, variation in the character's belief and desire modulated activity in bilateral TPJ, showing substantial overlap with TPJ voxels identified in the false belief/false photograph task. In contrast, neither the belief nor the desire factor modulated activity in ventral mPFC, though this brain region did show selective activity in the false belief/false photograph task. Participants in this study were clearly capable of engaging v-mPFC for mindreading, but did not appear to do so when they only had to store and use mental states to predict behaviour. In a second study, Hartwright et al. (2014) adapted the belief-desire task to reintroduce the need for a mindreading inference. In this task the character changed from trial to trial, there were prizes rather than foods, and the character's desire for the prize on offer was indicated through realistic photographs of faces that were smiling (positive desire), frowning (negative desire), or neutral (unknown desire). In the unknown desire condition participants had to make a mindreading inference about whether that character would want that prize. In this task, variation in the desire factor did modulate activity in v-mPFC, and this effect was driven by the unknown desire condition differing from the positive and negative desire conditions. These findings suggest that v-mPFC may have a distinctive role in mindreading inferences, and that the near ubiquity of activity in this region in studies of mindreading reflects the fact that most mindreading tasks entail mindreading inferences, and cannot distinguish activity due to these inferences from other component processes.

Of course, associating mindreading inferences with v-mPFC is just one step in understanding the cognitive basis of mindreading inferences and what role v-mPFC has in supporting these processes. As discussed earlier, mindreading inferences often involve complex integration of information from multiple sources under conditions of uncertainty in order to make a "best guess" about the target's mental states. The apparent simplicity of classic mindreading tasks, such as the "Sally-Anne" task, obscures the fact that it is only the pragrmatic context that suggests she must think it's either in the basket or the box: in fact Sally *could* think her ball is absolutely anywhere. Such uncertainty and context-sensitivity is much more apparent in more realistic mindreading situations (Apperly, 2010). The hypothesis that v-mPFC helps meet these functional requirements is supported by a study from Jenkins and Mitchell (2010) who independently varied whether a mindreading task required inferences about a character's mental states or their preferences, and whether those inferences were clearly warranted by the situation or were more uncertain and ambiguous. Whereas TPJ (and not mPFC) activity was sensitive to whether the inferences concerned mental states versus preferences, mPFC activity (and not TPJ) was sensitive to the level of uncertainty in the inference. Moreover, these findings converge with a broader literature that implicates v-mPFC in complex information integration and reasoning under uncertainty (e.g. Burgess, Dumontheil, & Gilbert, 2007).

A Future Prospect: Do Mindreading Brain Regions Represent What Others Are Thinking?

Since mindreading involves representing what other people are thinking (or feeling, or intending, etc.), and since mindreading recruits a reliable network of brain areas, it would be natural to suppose that one or all of these brain areas represents the thoughts of other people. Surprisingly, however, no evidence bears directly on this question, and in fact different theories about the "mindreading brain network" point towards different expectations. It is exciting that methods for decoding the informational content of neural activity (e.g. Haxby, Connolly, & Guntupalli, 2014; Kriegeskorte & Kievit, 2013) are opening up the possibility of directly testing such questions.

The idea that the "mindreading brain network" must be representing what other people are thinking seems a good hypothesis, and it clearly predicts that during a mindreading task TPJ and/or mPFC must be carrying information that distinguishes between instances in which Sally thinks her marble is in the basket, versus Sally thinks her marble is in the box, versus John thinks his marble is in the basket, etc. Put more operationally, if one trained a multivariate pattern classifier on patterns of activity in TPJ (for example) over a variety of instances in which an agent thinks an object is in a location, the classifier should be able to take new data from the same subject and distinguish trials on which Sally thinks the ball is in the basket from other combinations of information about agent-object-location combinations. Encouragingly, recent evidence suggests that category-level and even item-level information can be decoded from patterns of activity in TPJ and mPFC during memory retrieval (Kuhl & Chun, 2014; Zeithamova, Dominick, & Preston, 2012), suggesting that this question is tractable for suitably designed sets of Agent-Object-Location stimuli.

However, this outcome is far from a foregone conclusion. It is well-known that TPJ and mPFC are involved in attentional control, as well as mindreading (e.g. Burgess et al., 2007; Corbetta, Patel, & Schulan, 2008), and as discussed earlier there are good grounds for thinking that TPJ and mPFC may be specifically involved in controlling attention in order to maintain a distinction between information and processes related to self and other. This is compatible with the selective engagement of TPJ and mPFC in mindreading, but in no way entails that these regions represent the information about the agents, objects, locations, etc. over which they are exerting control; instead that information could be represented in participants' own primary semantic systems. Thus we do not yet have an answer to one of the most fundamental neuroscientific questions about mindreading: do mindreading brain regions represent information about mindreading?

Studies of mindreading have just begun to exploit the power of MVPA, successfully decoding broad types of social tasks and mental states from activation patterns in TPJ and/or mPFC (e.g. Koster-Hale, Saxe, Dungan, & Young, 2013; Tamir, Thornton, Contreras, & Mitchell, 2016). Extending this approach to examine how and when we represent the content of other minds not only addresses questions about how the brain supports mindreading. It also opens ways to tackle functional questions that have proved fiendishly difficult to address so far: Do perspectives of self and other recruit the same representational resources? Are self and other perspectives activated in series or in parallel? Do control processes, such as those associated with IFG, work to resolve competition between the content of self and other perspectives, or only competition between responses or judgements based on these perspectives. The role of IFG in inhibiting representational content during selective episodic memory retrieval (e.g. Wimber, Alink, Charest, Kriegeskorte, & Anderson, 2015) certainly makes it plausible that IFG also directly acts on the contents of self and other perspectives. In sum, MVPA offers the prospect of a rich interaction between cognitive and neuroscientific approaches through the common currency of "information".

Summary I have outlined a cognitive model of mindreading that is narrowly focused on processes directly involved in inferring, storing and using information about other people's mental states. A narrow focus makes it possible to think about the relationships between individual processing steps and their cognitive and neural bases, but of course it should not blind us to the fact that there is much more to mindreading than what I have discussed here. More ambitious and exhaustive models are very valuable but they face a daunting challenge in knowing where to stop. A good case can be made for including gaze processing, face recognition, moral and causal reasoning as part of mindreading (e.g. Schaafsma et al., 2015), However, following this logic, since I can imagine you thinking anything I can think for myself, there seems no principled limit on the information and processes on which I might need to draw, and so no straight-forward way of distinguishing between processes that are involved and not involved in mindreading. This is a deep issue with mindreading, but it should not stop us from building rich models of how mindreading is supported by a variety of cognitive and neural processes.

I hope I have also demonstrated that this is a two-way street, with results from neuroscientific studies informing cognitive theories just as much as the reverse. Relevant theories and methods must also interact. For example, it is important to recognize that subtractive neuroimaging designs optimized to detect domain-specific mindreading processes will tell us little about the nature of the processes involved, whereas designs that contrast different conditions within a mindreading task might tell you more about processes but little about their domain specificity. The rate of innovation in neuroscientific methods holds out great future promise for a cognitive neuroscience of mindreading, which will be maximized when combined with functional models of the cognitive processes involved.

References

Adolphs, R. (2009). The social brain: Neural basis of social knowledge. *Annual Review of Psychology, 60*, 693–716.

Apperly, I. A. (2008). Beyond simulation-theory and theory-theory: Why social cognitive neuroscience should use its own concepts to study "theory of mind". *Cognition, 107*, 266–283.

Apperly, I. A. (2010). *Mindreaders: The cognitive basis of "theory of mind"*. Hove, UK: Psychology Press/Taylor & Francis Group.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953–970.

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science, 17*(10), 841–844.

Apperly, I. A., Warren, F., Andrews, B. J., Grant, J., & Todd, S. (2011). Error patterns in the belief-desire reasoning of 3- to 5-year-olds recur in reaction times from 6 years to adulthood: Evidence for developmental continuity in theory of mind. *Child Development, 82*(5), 1691–1703.

Barwise, J., & Perry, J. (1983). *Situations and attitudes*. Cambridge, MA: MIT Press.

Brass, M., Bekkering, H., Wohlschläger, A., & Prinz, W. (2000). Compatibility between observed and executed finger movements: Comparing symbolic, spatial, and imitative cues. *Brain and Cognition, 44*(2), 124–143.

Burgess, P. W., Dumontheil, I., & Gilbert, S. J. (2007). The gateway hypothesis of rostral prefrontal cortex (area 10) function. *Trends in Cognitive Sciences, 11*(7), 290–298.

Cook, J. L. (2014). Task-relevance dependent gradients in medial prefrontal and temporoparietal cortices suggest solutions to paradoxes concerning self/other control. *Neuroscience and Biobehavioral Reviews, 42*, 298–302.

Corbetta, M., Patel, G., & Schulan, G. L. (2008). The reorienting system of the human brain: From environment to theory of mind. *Neuron, 58*, 306–324.

Davidson, D. (1990). The structure and content of truth. *Journal of Philosophy, 87*(6), 279–328.

Davies, M., & Stone, T. (Eds.). (1995). *Folk psychology: The theory of mind debate*. Oxford, England: Blackwell.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241.

Fauconnier, G. (1985). *Mental spaces: Aspects of meaning construction in natural language*. Cambridge, MA: MIT Press.

Fodor, J. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Friedman, O., & Leslie, A. M. (2004). Mechanisms of belief-desire reasoning: Inhibition and bias. *Psychological Science, 15*, 547–552.

Frith, C. D. (2007). The social brain? *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362*(1480), 671–678.

German, T., & Hehman, J. (2006). Representational and executive selection resources in 'theory of mind': Evidence from compromised belief-desire reasoning in old age. *Cognition, 101*(1), 129–152.

Hartwright, C. E., Apperly, I. A., & Hansen, P. C. (2012). Multiple roles for executive control in belief-desire reasoning: Distinct neural networks are recruited for self perspective inhibition and complexity of reasoning. *NeuroImage, 61*(4), 921–930.

Hartwright, C., Apperly, I. A., & Hansen, P. C. (2014). Representation, Control or Reasoning? Distinct Functions for Theory of Mind within the Medial Prefrontal Cortex. *Journal of Cognitive Neuroscience, 26*(4), 683–698.

Hartwright, C. E., Hardwick, R., Apperly, I. A., & Hansen, P. (2016). Structural morphology in resting state networks predict the effect of theta burst stimulation in false belief reasoning. *Human Brain Mapping, 37*, 3502–3514.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience, 37*, 435–456.

Heyes, C. (2014). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science, 9*, 131–143.

Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking.* Cambridge, MA: Harvard University Press.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences, 20*(8), 589–604.

Jenkins, A. C., & Mitchell, J. P. (2010). Mentalizing under uncertainty: dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cerebral Cortex, 20*(2), 404–410.

Keysar, B., Lin, S., Barr, D.J., (2003). Limits on theory of mind use in adults. *Cognition 89*, 25–41.

Koster-Hale, J., & Saxe, R. (2013). Functional neuroimaging of theory of mind. In S. Baron-Cohen, M. Lombardo, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental social neuroscience* (pp. 132–163). Oxford, England: Oxford University Press.

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Science, 110*, 5648–5653.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science, 330*, 1830–1834.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences, 17*, 401–412.

Kuhl, B. A., & Chun, M. M. (2014). Successful remembering elicits event-specific activity patterns in lateral parietal cortex. *The Journal of Neuroscience, 34*(23), 8051–8060.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind". *Psychological Review, 94*, 412–426.

Leslie, A. M. (2005). Developmental parallels in understanding minds and bodies. *Trends in Cognitive Sciences, 9*(10), 459–462.

Low, J., Apperly, I. A., Butterfill, S. A., & Rakoczy, H. (2016). Cognitive architecture of belief reasoning in children and adults: A primer on the two-systems account. *Child Development Perspectives, 10*(3), 184–189.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W.H. Freeman.

McCleery, J. P., Surtees, A. D., Graham, K. A., Richards, J. E., & Apperly, I. A. (2011). The neural and cognitive time course of theory of mind. *The Journal of Neuroscience, 31*(36), 12849–12854.

Perner, J., & Leekam, S. (2008). The curious incident of the photo that was accused of being false: Issues of domain specificity in development, autism, and brain imaging. *The Quarterly Journal of Experimental Psychology, 61*(1), 76–89.

Phillips, J., Ong, D. C., Surtees, A. D., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A second look at automatic theory of mind: Reconsidering Kovács, Téglás, and Endress (2010). *Psychological Science, 26*(9), 1353–1367.

Qureshi, A., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective-selection, not Level-1 visual perspective-calculation: Evidence from a dual-task study of adults. *Cognition, 117*(2), 230–236.

Royzman, E. B., Cassidy, K. W., & Baron, J. (2003). "I know, you know": Epistemic egocentrism in children and adults. *Review of General Psychology, 7*(1), 38–65.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance, 36*, 1255–1266.

Samson, D., Apperly, I. A., Kathirgamanathan, U., & Humphreys, G. W. (2005). Seeing it my way: A case of selective deficit in inhibiting self-perspective. *Brain, 128*, 1102–1111.

Samson, D., Houthuys, S., & Humphreys, G. W. (2015). Self-perspective inhibition deficits cannot be explained by general executive control difficulties. *Cortex, 70*, 189–201.

Santiesteban, I., Banissy, M. J., Catmur, C., & Bird, G. (2012). Enhancing social ability by stimulating right temporoparietal junction. *Current Biology, 22*, 2274–2277.

Santiesteban, I., White, S., Cook, J., Gilbert, S. J., Heyes, C., & Bird, G. (2012). Training social cognition: From imitation to theory of mind. *Cognition, 122*, 228–235.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *Neuroimage, 19*, 1835–1842.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences, 19*(2), 65–72.

Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience, 10*, 640–656.

Stuhlmüller, A., & Goodman, N. D. (2014). Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Cognitive Systems Research, 28*, 80–99.

Surtees, A., Samson, D., & Apperly, I. A. (2016). Unintentional perspective-taking calculates whether something is seen, but not how it is seen. *Cognition, 146*, 97–105.

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Science, USA, 113*(1), 194–199.

van der Meer, L., Groenewold, N.A., Nolen, W.A., Pijnenborg, M., Aleman, A., (2011). Inhibit yourself and understand the other: Neural basis of distinct processes underlying Theory of Mind. *NeuroImage 56*, 2364–2374.

van der Wel, R. P., Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition, 130*(1), 128–133.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happe, F., Falkai, P., … Zilles, K. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *NeuroImage, 14*(1).

Wagner, A. D., Maril, A., Bjork, R. A., & Schacter, D. L. (2001). Prefrontal contributions to executive control: fMRI evidence for functional distinctions within lateral prefrontal cortex. *Neuroimage, 14*(6), 1337–1347.

Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories. *Nature Neuroscience, 18*, 582–589.

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron, 75*, 168–179.

The Neural Basis and Representation of Social Attributions

Frank Van Overwalle and Elien Heleven

Introduction

When maneuvering through our social world and meeting other people, it is crucial to understand their behaviors and minds. The capacity to understand another person's emotions, intentions, beliefs, and personality traits, based on observed or communicated behavior, is termed social cognition. During the last decade, neuroscience has greatly increased our insights about social cognition by studying the neural correlates of its underlying processes and representations. Two main cortical networks have been identified (Fig. 1): The *mirror* network recruited when we observe the actions of other persons (i.e., "body" reading) and which is part of a larger *sensorimotor* network (Yeo et al., 2011), and the *mentalizing* network activated when we imagine the mental state of another person (i.e., "mind" reading; for reviews, see Schurz, Radua, Aichhorn, Richlan, & Perner, 2014; Van Overwalle, 2009) and which is part of the larger *default* network (Raichle et al., 2001). Stated differently, mirroring reflects lower-level processes of immediate motion and action perception (e.g., biological face, arm and leg moves, and gestures), while mentalizing involves higher-level inference about non-observable mental entities such as intentions, beliefs, and traits. These two networks of social cognition operate largely independently from each other (Van Overwalle & Baetens, 2009). Higher-level mentalizing on intentions, beliefs, and traits is a capacity that is most developed and sophisticated in humans compared to animals (e.g., primates), and this chapter focuses on the neural basis of this mentalizing processes in social cognition.

F. Van Overwalle (✉) · E. Heleven
Department of Psychology & Center of Neuroscience, Vrije Universiteit Brussel,
Brussel, Belgium
e-mail: Frank.VanOverwalle@vub.ac.be

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_19

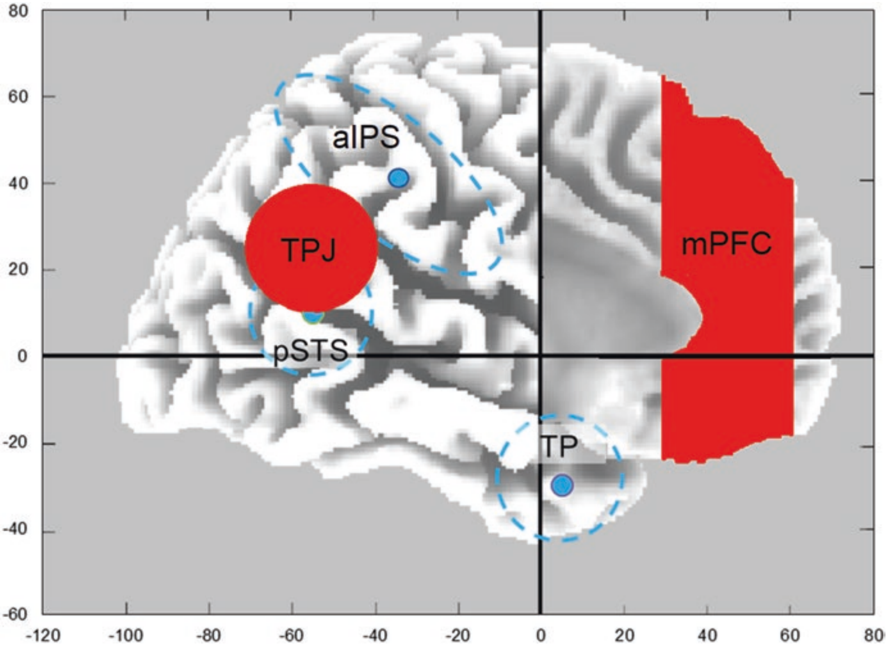


Fig. 1 Lateral and medial side of the human brain (left and right panel, respectively), with a schematic indication of the core mentalizing areas (in red) and other socially relevant areas (in blue) together with their primary social function (between parentheses): *TPJ* temporo-parietal junction (temporary judgments such as other people's intentions and beliefs), *mPFC* medial prefrontal cortex (enduring judgments such as traits and preferences), *pSTS* posterior superior temporal sulcus (biological movement; part of the mirror network), *aIPS* anterior intraparietal sulcus (biological movement in the context of an object; part of the mirror network), *TP* temporal pole (semantic memory of social contexts and scripts). Not shown is the precuneus (PC; subserving autobiographic memory; part of the mentalizing network) which lies at approximately the same position as the TPJ on the medial side of the brain

Do I Have Control over My Social Judgments? Implicit and Explicit Social Mentalizing Driven by a Shared Brain Network

Contrary to the old idea that social attributions require a lot of explicit or deliberate mental elaboration, behavioral research in the 1980s (Winter & Uleman, 1984) documented that social attributions, including trait inferences, are often made implicitly, spontaneously, and automatically, without awareness or control about the inference process. An increasing number of studies demonstrate that representing other agents' beliefs is an implicit capacity acquired early at 7 months of age (Kovacs, Téglás, & Endress, 2010) and implicitly sustained during adulthood (Schneider, Bayliss, Becker, & Dux, 2012), although it requires some minimal executive resources (Qureshi, Apperly, & Samson, 2010; Schneider, Lam, Bayliss, &

Dux, 2012). Although such quick judgments are relatively correct (Letzring, Wells, & Funder, 2006), a fascinating question is how much they differ from explicit attributions? Do they rely on different processes and neural correlates? Although some theorists initially proposed distinct brain areas (e.g., Forbes & Grafman, 2013; Lieberman, 2007; Satpute & Lieberman, 2006), recent research at our lab and by other researchers demonstrated that that implicit and explicit person inferences do not rely on strictly distinct neural processes or substrates (e.g., Kestemont, Vandekerckhove, Ma, Van Hoeck, & Van Overwalle, 2013; Ma, Vandekerckhove, Van Overwalle, Seurinck, & Fias, 2011).

The idea of a great divide between implicit and explicit processes stems from prior dual-process theories (for reviews see Evans, 2008; Evans & Stanovich, 2013), which make a distinction between processes that are *implicit* (also termed unconscious, automatic, spontaneous, experiential, heuristic, intuitive, impulsive, and reflexive) and *explicit* (also termed conscious, controlled, rational, systematic, analytical, and reflective; Chaicken, 1980; Epstein, 1994; Lieberman, 2007; Schneider & Shiffrin, 1977; Strack & Deutsch, 2004). There are indeed neural network that supports mainly implicit processes, such as subcortical mechanisms located in the amygdala and other limbic structures which elicit primitive affective reactions (e.g., rapid impressions of a face; Forbes, Cox, Schmader, & Ryan, 2012; Todorov, Baron, & Oosterhof, 2008; Todorov, Said, Engell, & Oosterhof, 2008), and mirror-like neural networks which support implicit understanding of non-verbal actions of humans (Iacoboni, 2009; Van Overwalle & Baetens, 2009).

In contrast, higher-level mentalizing brain areas subserve computations that are neither exclusively implicit nor explicit (Forbes & Grafman, 2013; Keysers & Gazzola, 2007; Satpute & Lieberman, 2006). The neural network responsible for mentalizing is illustrated in Fig. 1 (Yeo et al., 2011). Each core area in this mentalizing network is recruited for a distinct computation and specific input and does not depend on an implicit or explicit mode of processing. Van Overwalle (2009) argued that the *temporo-parietal junction* (TPJ) seems responsible for judgments on temporary beliefs and intentions, while the *medial prefrontal cortex* (mPFC) seems involved in enduring trait inferences and other stable characteristics (see for reviews, Amodio & Frith, 2006; Bzdok et al., 2012; Denny, Kober, Wager, & Ochsner, 2012; Lombardo, Chakrabarti, Bullmore, & Baron-Cohen, 2011; Van Overwalle, 2009). Figure 1 shows the location of these two key mentalizing networks, together with other adjacent areas of the mirror network and their proposed functionality (pSTS & aIPS).

Implicit and Explicit Mentalizing Share Early Timing and Core Brain Areas

To uncover the neural correlates of implicit and explicit social thinking, research in our lab typically used a straightforward approach. Participants were given behavioral descriptions which were preselected so that they would spontaneously elicit specific social attributions such as goals, beliefs, or traits. One half of the participants received the instruction to read the material for understanding, while another half were explicitly instructed to make this specific social attribution. By using two groups, leakage of explicit instructions into the implicit condition was completely avoided. The implicit instruction to simply read the material seems methodological preferable for neuroimaging research, because it allows to investigate the process of interest directly, without confounding manipulations and their concurrent brain activations (e.g., cognitive load, indirect memory measures). Nevertheless, memory measures were often taken after the experiment to make sure that social attributions were made to the same degree under both implicit and explicit instructions. These experiments demonstrated that there is a common underlying mentalizing network that is relatively blind to the implicit (or spontaneous) versus explicit (or deliberate) nature of the attribution, and that seems more sensitive to the content of the attribution.

In a first set of experiments using electroencephalogram (EEG) measures, Van Overwalle and colleagues documented that the neural timing of the onset of an early social inference is almost identical under implicit or explicit processing. They measured the onset by presenting inconsistent information (e.g., a friendly person gives a "slap") and looking at the EEG signal reflecting this inconsistency. These studies demonstrated that irrespective of the implicit or explicit instruction, goal attributions were made after about 250 ms (Van der Cruyssen, Van Duynslaeger, Cortoos, & Van Overwalle, 2009) and that trait attributions occurred at about 600 ms (Van Duynslaeger, Van Overwalle, & Verstraeten, 2007). These results seem to provide support for a core single-system account with an identical onset at the beginning of mentalizing. That attributions on goals were faster than on traits is consistent with the proposition by Van Overwalle (2009) that goals involve a quick evaluation of transient mental contents, that pertains to the here-and-now by the TPJ, while traits reflect slowly generated abstractions by the mPFC, extracted from behaviors identified in the TPJ (see also Ma, Vandekerckhove, Van Hoeck, & Van Overwalle, 2012; Van Overwalle, Van Duynslaeger, Coomans, & Timmermans, 2011). Source localization of the EEG waves suggested that the core mentalizing areas (TPJ and mPFC) were most strongly recruited (using Loreta, Pascual-Marqui, 1999; Pascual-Marqui, Michel, & Lehmann, 1994).

In a second set of experiments using functional magnetic resonance imaging (fMRI), Van Overwalle and colleagues conducted a series of studies to explore the overlap in brain areas involved in explicit and implicit mentalizing. They used a very similar experimental design as described before, with two groups receiving implicit reading or explicit attributional instructions. The results showed significant overlap in mentalizing activity after implicit and explicit instructions. An fMRI study on trait inferences revealed common activation in the mentalizing network (Fig. 2a; Ma et al., 2011). Importantly, there were also differences between instructions. Implicit trait inferences significantly recruited only the core mentalizing areas of the TPJ and mPFC, whereas explicit trait attributions additionally recruited other brain areas involved in mentalizing, including the precuneus (responsible for autobiographic memory and scene construction) and posterior part of the superior temporal sulcus (pSTS; involved in detecting biological motion). Analogous findings were reported by Rameson, Satpute, and Lieberman (2010) for implicit and explicit self-descriptions.

Another fMRI study on the attribution of causes of events to persons or situations (Kestemont et al., 2013) showed an overlap for implicit and explicit instructions in the bilateral TPJ and pSTS (Fig. 2b). Again, there were also differences. Only implicit inferences increased the activation of the mPFC, suggesting a tendency to

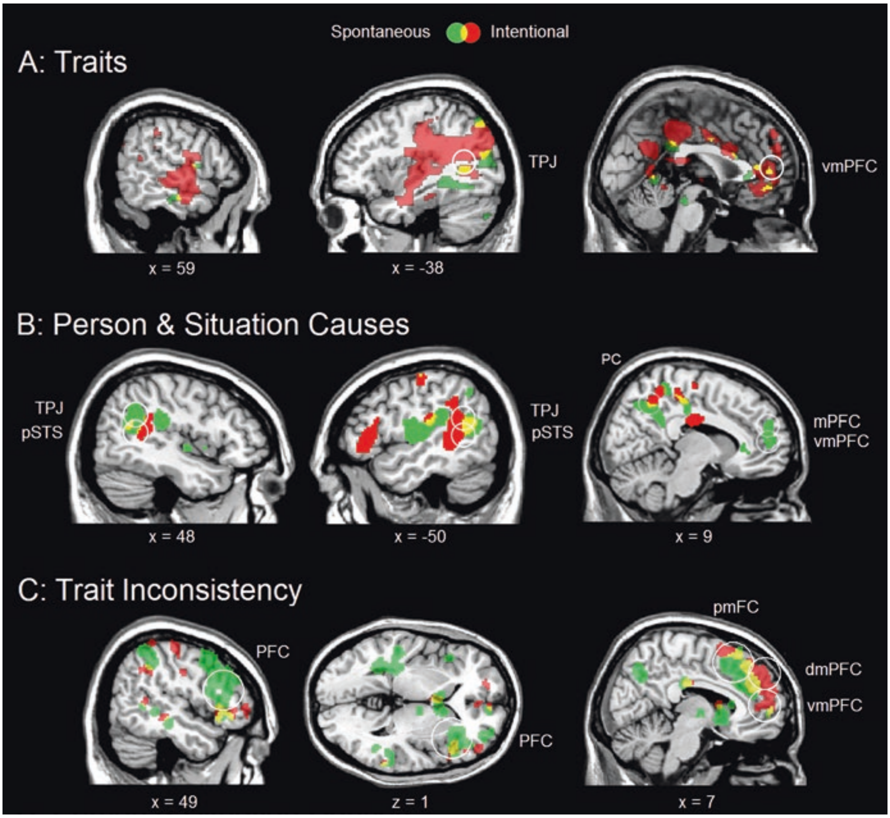


Fig. 2 fMRI activations given various social inferences under spontaneous (green) and intentional (red) instructions, and their overlap (yellow). (**A**) Consistent trait > irrelevant trait contrast from Ma et al. (2011) with significant overlap/conjunction in the left TPJ (MNI coordinates −58 −58 32). (**B**) Person Cause > Baseline & Situation Cause > Baseline conjunction from Kestemont et al. (2013) with significant conjunction in the bilateral TPJ and pSTS (respective MNI coordinates 46 –56 20; −50 −54 18; 52 –56 14; −50 −56 16). (**C**) Inconsistent trait > Consistent trait contrast from Ma, Vandekerckhove, Baetens, et al. (2012), Ma, Vandekerckhove, Van Hoeck, and Van Overwalle (2012) with significant conjunction in the mPFC (MNI coordinates 4 42 32). In all analyses, whole-brain activation was thresholded at $p < 0.005$ (uncorrected) with at least 10 voxels. Circles indicate regions of interest with significant activation after FDR correction at $p < 0.10$. *vmPFC* ventral part of the mPFC, *dmPFC* dorsal part of the mPFC, *pmFC* posterior frontal cortex, *PFC* lateral prefrontal cortex

make dispositional trait attributions to the person, known as the *fundamental attribution bias* (Ross, 1977). This biased activation of the mPFC was absent under explicit instructions, consistent with decreased biased processing documented in behavioral research (Gilbert & Malone, 1995) and neuroimaging (Brosch, Schiller, Mojdehbakhsh, Uleman, & Phelps, 2013).

A last fMRI study involved sets of trait-implying sentences which were interspersed with sentences that implied an inconsistent trait (Ma, Vandekerckhove, Baetens, et al., 2012). This study revealed a significant overlap in the dorsal part of the mPFC (Fig. 2c). Like in the previous trait study, some brain areas were more active only under explicit instructions, including the left TPJ and pSTS (biological motion) and the precuneus (autobiographic memory and scene construction). Interestingly, both instructions also revealed a significant overlap in the posterior medial frontal cortex (pmFC, including the dorsal part of the anterior cingulate cortex—dACC) and the right PFC. These latter two areas are part of a domain-general conflict monitoring network (Botvinick, Cohen, & Carter, 2004) that detects and resolves conflicts between multiple or conflicting inputs.

Implicit and Explicit Mentalizing as Iterative Reprocessing

Taken together, these neuroscientific data from our lab suggest that implicit and explicit mentalizing share the same early timing and the same core brain areas, but also that explicit attributions may lead to a modulation in some additional brain areas, perhaps reflecting a correction or an enrichment. This is broadly in line with the proposal by Tamir and Mitchell (2010, 2012) that perceivers extract social inferences from an initial starting point that quickly comes to mind, and then customize it by adjusting away from this anchor. However, unlike this proposal, we do not suggest that this initial anchor is necessarily based on knowledge about the self, and that adjustments are made for persons who are less similar to the self. Instead, we speculate that implicit information might be enriched under explicit instructions (1) by retrieving similar behaviors from the past and imaging more vividly social cues on human action and movement leading to more activation in the middle temporal lobe (Ma, Vandekerckhove, Baetens, et al., 2012, Ma et al., 2011), (2) by relying on autobiographical memories and social background or scenes in the precuneus (Ma, Vandekerckhove, Baetens, et al., 2012), or (3) by taking in more situational information so that a biased trait attributions putting the person on the foreground is avoided, leading to less activation in the mPFC (Kestemont et al., 2013). Evidently, these speculations need to be tested in future research.

Nonetheless, the present data are best explained by an *iterative reprocessing* model (Cunningham & Zelazo, 2007). According to this model, processing occurs on a continuum from relatively implicit to relatively explicit. Increased explicit processing is possible through additional reprocessing cycles, which enable more explicit elaboration of information along a wider and richer range of contexts and constraints, retrieving input from increasingly more brain structures. With each iteration cycle, information is passed back and forth. Thus, inferences based on one or a few cycles are relatively implicit and crude intuitions, leaving an early mark in the EEGs and activation in restricted core brain areas. Later on, inferences based on additional iterations and computations are increasingly rich, balanced and relatively explicit, leaving a broader trace of activation in extended brain areas.

To sum up, social neuroscience demonstrated that understanding of another persons' mind involves both implicit and explicit processes located in the mentalizing network (e.g., Keysers & Gazzola, 2007), contradicting old ideas that that these processes are driven by entirely different underlying brain systems. Under implicit and explicit processing instructions, there was a shared early timing (EEG studies) and shared brain activity (fMRI studies) during goal, causal and trait attributions, pointing to a single core system of mentalizing (Van Overwalle, 2009). Neuroimaging research from other labs confirms that other mentalizing tasks such as false beliefs also recruit a common set of key mentalizing brain areas under implicit and explicit processing (Kovács, Kühn, Gergely, Csibra, & Brass, 2014; Naughtin et al., 2017; Schneider, Slaughter, Becker, & Dux, 2014). The present evidence further suggests that there is an implicit default core process that allows observers to make quick social mentalizing inferences, presumably based on current information and preexisting learned social knowledge. This implicit core process is subserved by the TPJ and mPFC. Subsequent reprocessing cycles allow to take in more and richer information from other brain areas which enable observers to verify and flexibly control their original rapid intuition.

How Do I Access My Social Knowledge? The Role of the TPJ in Reorienting to Social Content

The previous section demonstrated that social neuroscience during the last two decades made enormous strides in identifying the neural basis of social processes. It is now becoming increasingly clear that the two key areas of the social mentalizing system, the TPJ and mPFC (see meta-analyses by Decety & Lamm, 2007; Schurz et al., 2014; Van Overwalle & Baetens, 2009) are involved in different processes.

Let us begin with the TPJ. Research has shown that this area is not only part of the mentalizing network, but also of the ventral attention network involved in attention allocation (see review by Cabeza, Ciaramelli, & Moscovitch, 2012). Unexpected stimuli in our environment are captured by a ventral attention network including the right TPJ, which reorients attention to these salient stimuli. The dual role in mentalizing and attention has been confirmed in meta-analyses of neuroimaging research (Decety & Lamm, 2007; Krall et al., 2014; Van Overwalle & Baetens, 2009), as well as in research on the same participants (Mitchell, 2008; Scholz, Triantafyllou, Whitfield-Gabrieli, Brown, & Saxe, 2009). Note that this function differs markedly from adjacent mirror areas in Fig. 1 which rely on the direct observation of biological movement, such as the pSTS (identifying biological movement) and aIPS (understanding biological movement in the context of a manipulated object).

This functional overlap in the right TPJ has important theoretical implications on the role of the TPJ in mentalizing. Some theorists put forward an underlying shared process (Krall et al., 2014). Among the proponents of this view, Van Overwalle (2009) suggested that attributions of intentionality and mental state require a "where-to" shifting function that requires attention to relevant information, and is also evident in attention reorientation. Attention and action directed toward a specific entity often express the intention to reach that entity. Along somewhat different lines, Cabeza et al. (2012) proposed a theoretical model in which attention can be directed not only to unexpected external stimuli (cf. basic attention reorientation), but also to one's internal memory. This can explain the role of the TPJ in belief reasoning. Indeed, understanding that another person holds mental beliefs requires turning one's attention to memories on the person's recent behavior in a given context (see also further on "false" beliefs). It may also require directing one's attention to basic social knowledge, in order to extract implicitly or explicitly some basic social attributions (e.g., traits) from this behavior.

To study spatial reorientation, researchers typically use Posner's (1980) *cuing task*. In this task, a cue (e.g., an arrow) indicates the location of the upcoming target stimulus in most trials correctly (valid trials), while in other trials the cue indicates the incorrect location (invalid trials). Invalid trials require participants to disengage from their current visual focus suggested by the cue, and to reorient one's attention to another part of the visual field where the target stimulus appears. Research has shown that invalid trials lead to higher activation of the right TPJ (see meta-analyses by Decety & Lamm, 2007; Van Overwalle & Baetens, 2009).

To study social mentalizing, researchers often used *false beliefs*, which is the understanding of another person's beliefs that involve an element of false (outdated) knowledge (e.g., Saxe & Kanwisher, 2003). For instance, when an agent is unaware of changes in an object's location during his or her absence (e.g., when someone else took a toy or candy away), we need to understand that the agent is convinced of his or her original belief (e.g., thinking that the toy or candy is still in its original location), although this belief is false with respect to current reality, hence the term "false" beliefs. Understanding false beliefs is a key process in mentalizing, because it requires participants to direct their attention internally to false mental states, while for true beliefs they can simply observe what is out in reality. Evidence shows that false beliefs lead to increased engagement of the right TPJ in comparison with true beliefs (see meta-analyses by Schurz et al., 2014; Van Overwalle, Van den Eede, Baetens, & Vandekerckhove, 2009).

Does the increased activation of the right TPJ under attention reorientation and social attribution demonstrate that a common core process underlies both? A recent meta-analysis by Krall et al. (2014) seems to confirm a joint overarching account of the TPJ. These authors found that the anterior TPJ is a common area supporting reorientation of attention and false belief, while the posterior TPJ was found to support only social belief processes. A limitation, however, is that earlier research used different modalities, measuring attention reorientation with Posner's visual task and false belief mentalizing with verbal stories. Perhaps the unique activation of the posterior TPJ for false beliefs was due to its verbal nature, rather than something specific about social mentalizing.

To resolve this limitation, in a recent study in our lab, Özdem and coworkers (Özdem, Brass, Van der Cruyssen, & Van Overwalle, 2017) kept the visual/spatial modality of both tasks alike, by using a novel false belief task that was visually and spatially quite similar to the Posner reorientation task. Specifically, in the Posner task, an arrow was used as cue after which a target stimulus (a black circle) appeared on the left or right side of a window (Fig. 3a left). In the mentalizing task, the same arrow was now an agent, faced toward the window, who might stay or leave, and consequently might (or might not) witness the change of location of the black circle to left or right side of the window (Fig. 3a right). The results clearly showed an overlap in TPJ activation between the spatial versions of the reorientation and mentalizing tasks (Fig. 3b), with additional areas uniquely related to attention reorientation and belief mentalizing. This is in line with the attention orientation account proposed by Cabeza et al. (2012), which suggests that the TPJ has an overarching common attention function and that various subregions of the TPJ mediate different aspects of related subprocesses through connectivity with different inputs (e.g., spatial location of external stimuli in the Posner task; internal memory in a false belief task). We speculate that unexpected false stories involve high-level disruption of attention, while the classic Posner task involves low-level attention reorientation (see also Krall et al., 2014).

To further test that the TPJ is involved in attention orientation towards internal memories about false beliefs, another study (Özdem, Brass, Schippers, Van der Cruyssen, & Van Overwalle, 2018) investigated the effect of two agents holding both false and/or true beliefs, rather than a single agent as in prior research. Participants saw animated stories with two smurfs witnessing (or not) a back circle changing its position on the screen, and thus holding true (or false) beliefs. Afterwards, they had to take the perspective of one of the smurfs or the self. Consistent with the idea that the TPJ is involved in attention to others' false beliefs held in memory, the results showed that when taking the perspective of one of the smurfs, TPJ activation linearly increased the more smurfs held a false belief.

Taken together, in line with the attention orientation account by Cabeza et al. (2012), we proposed that directing one's attention to internal memory and extracting information from it can explain the role of the TPJ in belief reasoning. To get grip on another person's beliefs requires turning one's attention to memories about this person's recent behavior in a given context (e.g., was she present or not when her boyfriend told a joke about here) and to recall it. On the basis of general social knowledge, it is possible to make appropriate social attributions (e.g., she would feel angry when he made fun of her). The key question now is: where is this high-level social knowledge on persons and traits stored in the brain? This is the topic of the next section.

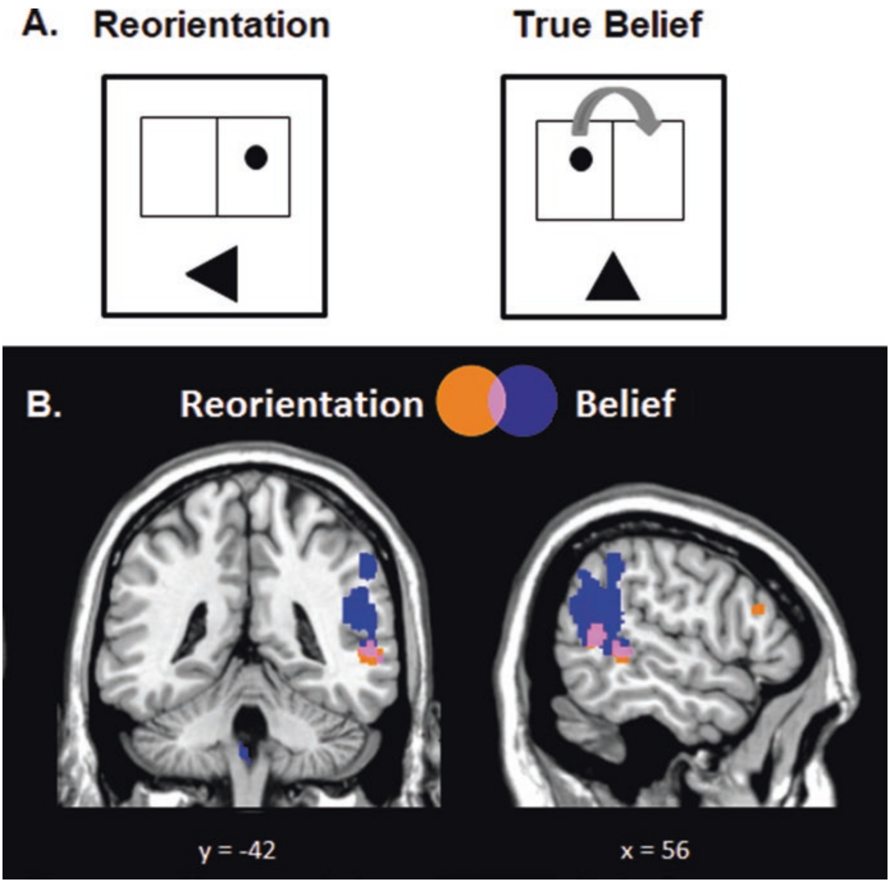


Fig. 3 (**a**) **Left**: An invalid orientation trial with an arrow cueing towards the left window but where the black target circle appears in the (opposite) right window (valid trials are similar, with the arrow curing in the correct direction). **Right**: A true belief trial in which the arrow faces the window and observes the target circle jumping from left to right (false belief trials are similar, but with the arrow being temporary absent). **(b**) The overlap between spatial reorientation (Invalid > Valid) and spatial belief (False Belief > True Belief). The clusters are whole-brain thresholded at $p < 0.001$ (uncorrected) with at least 10 voxels, and with significant activation after FWE correction at $p < 0$ .05 for regions of interest at the anterior TPJ (From Özdem et al., 2017)

Where Is Social Knowledge Stored? Social Representations in the Brain

A search for the neural representation of social knowledge is not straightforward, because it needs to avoid confounds such as processes that run in parallel and are thus difficult to disambiguate, although they are not involved in the critical process of interest. For instance, in order to infer a personality trait implied by someone's behavior (e.g., giving a slap), we need to understand the behavior based on the prior *intentions* of the agent (e.g., bringing harm/back to consciousness), and we often may experience *emotional* consequences (e.g., anger/relief) and *behavioral* tendencies (e.g., punish/reward someone). To avoid these parallel confounding processes, research in our lab used fMRI repetition suppression. This paradigm is based on the idea that whenever information is processed, this leaves traces of activation in the neuronal population where this information is stored in the brain. When the same information is processed again, this neuronal population will immediately "recognize" it and process it much more efficiently (Grill-Spector, Henson, & Martin, 2006; Wood & Grafman, 2003). This leads to reduced brain activation which is termed *repetition suppression* (Grill-Spector et al., 2006), but only in the area where the information is stored. Thus, in contrast to a typical analysis of activation caused by the whole information stream, repetition suppression focuses exclusively on the area that reflects the representation of knowledge and that shows suppression after repeating the information.

Although repetition suppression was already applied previously for the identification of neural representation of low-level social visual information (e.g., faces; Avidan, Hasson, Hendler, Zohary, & Malach, 2002; Henson, 2000; for a meta-analysis: Kim, 2017), it was only quite recently applied to identify higher-level representations such as those for action observation (Ramsey & Hamilton, 2010a, 2010b) and action word reading (Yee, Drucker, & Thompson-Schill, 2010).

Several studies recently showed that we also hold high-level knowledge of persons and their personality traits (e.g., Heleven & Van Overwalle, 2016; Ma et al., 2014). We recently summarized these fMRI suppression studies, mainly from our lab, in a meta-analysis (Heleven & Van Overwalle, 2018a). The results revealed that knowledge on traits and agents is represented in the ventral part in the mPFC (vmPFC) which are partly overlapping across traits and agents (Fig. 4). This area represents trait and agent knowledge about a plethora of people, including familiar and close persons such as oneself (Heleven & Van Overwalle, 2019), friends and family (Heleven & Van Overwalle, 2016), as well as unfamiliar people (Heleven, Boukhlal, & Van Overwalle, 2018), regardless of how well we know them (Heleven & Van Overwalle, 2018b).

Current research in our lab now also looks into representations of social categories or groups and the stereotypical traits we attribute about them. In a recent study we investigated the representation of professional groups such as nurses, agents, and so on (Delplanque, Heleven, & Van Overwalle, 2019). The findings of this suppression study confirm the mPFC as location of stereotypes, as one would expect given that these stereotypes often involve traits and other stable characteristics of group members. However, surprisingly, the professional groups themselves are represented in the posterior cingulate, an area nearby the precuneus which is also part of the mentalizing network. Given that this area is associated with scene construction and imagination, the results seem to suggest that we see groups rather as part of our social background and context, and less as a prominent agent.

These findings lend support to theories of person impression that see the ventral mPFC as a memory pool of person-related social information, including their traits and stereotypes. According to these theories, social information in the ventral mPFC

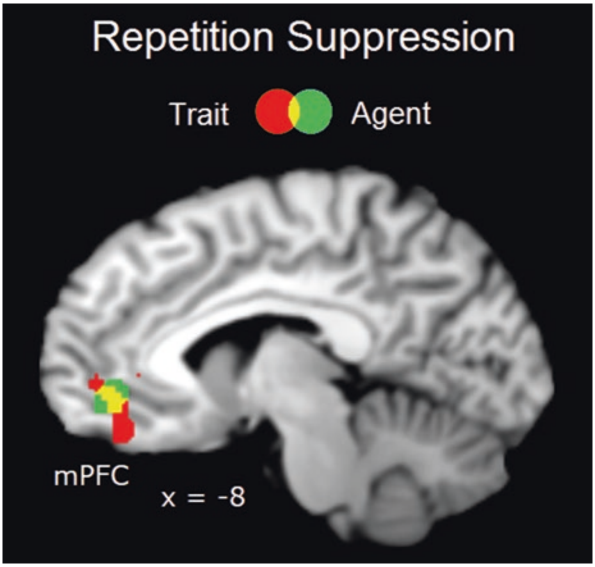


Fig. 4 Knowledge representations on traits and agents in the ventral part in the mPFC which are partly overlapping. Meta-analytic ALE activation maps (Eickhoff, Laird, & Grefkes, 2009; Laird et al., 2005) of repetition suppression overlaid on the Colin brain template. Shown are ALE clusters with $p < 0.05$, FDR corrected and volume > 200 mm^3. (From Heleven and Van Overwalle 2018a)

provides an initial anchor or estimate for social judgments, which is used to "simulate" or "project" it to other individuals whereby the dorsal mPFC subserves the adjustment of this initial anchor to each individual person (Tamir & Mitchell, 2010, 2012). Note, however, that many of these earlier theories viewed the ventral mPFC as a memory reservoir for traits and mental states mainly of the self, in which the self is used an initial template or anchor to judge others (Mitchell, 2009; Northoff et al., 2004). In contrast, the present results extend this idea and suggest that this area might have a more general function in the representation of agent and trait-related information per se, irrespective of the self (see also Welborn & Lieberman, 2015). It may thus represent a greater reservoir of trait knowledge about a plethora of people, including oneself, friends and family, as well as generic information applicable to anyone, including unfamiliar people. This greater pool of trait-related information, rather than only related to the self, might be used to judge others. For example, one might judge another person as very similar in character to oneself, but also in comparison with our mother, our best friend, a typical teacher, and so on.

Our finding of trait and agent representations in the ventral mPFC is also in line with the social psychology literature that conceives traits as "abstract instances of goal-directed behaviors" (see also Read, 1987; Read et al., 1990; Reeder, 2009; Reeder et al., 2004). Neuroimaging studies confirm this hierarchical relationship by demonstrating that trait inferences involve high-level abstractions of agent characteristics based on lower-level behaviors (Baetens, Ma, Steen, & Van Overwalle, 2014; Gilead, Liberman, & Maril, 2013). There is indeed evidence showing that the mPFC interacts with and forms links to other mentalizing areas such as the TPJ (Van Overwalle, Van de Steen, & Mariën, 2018; see Fig. 5) in order to enable lower level information retrieval and integration.

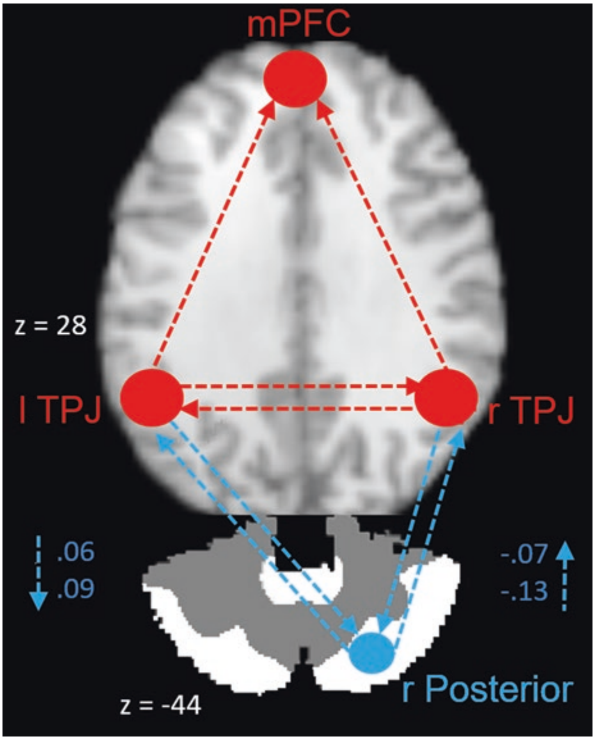


Fig. 5 Effective closed-loop connectivity in social mentalizing between the cerebellum and the cerebral cortex, simplified from Van Overwalle et al. (2018). The cerebellum (bottom) shows the mentalizing network colored in white and the other networks in gray. The strength of the connections is summarized on the left (from cortex to cerebellum; all >0) and right (from cerebellum to cortex, all <0). The major mentalizing hub in the cerebellum is located in the right posterior part (in blue). *mPFC* medial prefrontal cortex, *TPJ* temporo-parietal junction, *l* left, *r* right

This view of traits and agents as abstract instances of social cognition is in line with an approach which conceives the mPFC as an amodal hub or convergence zone of social information processing and social representations (Forbes & Grafman, 2010; Harada, Li, & Chiao, 2010; Patterson, Nestor, & Rogers, 2007; Woollams, 2012). Consistent with this approach, we interpret the representations for traits and agents in the mPFC as strongly interlinked and high-level abstract summary representations in a social hub that integrates information on actions and behaviors at lower levels of the hierarchy (see also Heleven & Van Overwalle, 2016; Krueger, Barbey, & Grafman, 2009).

The view of a social hub runs parallel with current theories on the neural basis of general semantic knowledge. One of these more recent theories, the "hub-and-spoke" model (Ralph, Jefferies, Patterson, & Rogers, 2017) proposes that verbal and non-verbal experiences provide the basic ingredients for constructing concepts and are encoded in modality-specific cortices distributed across the brain (the "spokes"). These provide the foundation for everyday behavior such as spreading jam on bread, which require the necessary knowledge on the qualities of objects and the deployment of appropriate movements. Crucially, all these modality-specific sources of information interact, at least in part, through a single transmodal semantic hub. Ralph et al. (2017) argued that this amodal hub for semantic cognition is situated in the bilateral anterior temporal lobes. We agree that social experiences arising from different modalities provide a basis for constructing social concepts and are encoded in modality-specific cortices as explained above. However, we diverge from this semantic theory by proposing that social concepts interact through the modulation of an additional, social hub specialized in mentalizing inferences and located in the mPFC. In other words, major semantic approaches (Binder, Desai, Graves, & Conant, 2009) have largely neglected the special place social mentalizing has in our cognition and neural processes (see also Van Overwalle, 2011). These semantic theories need to be extended to account also for a specialized amodal social hub that binds social concepts together at an abstract level.

The Right Sequence of Actions: The Cerebellum and Social Mentalizing

To this date, social neuroscience predominantly focused on the cerebral cortex and the role of the mirror and mentalizing network (Schurz et al., 2014; Van Overwalle & Baetens, 2009). However, some recent findings from our lab increased the interest of the scientific community for the role of the cerebellum in social cognition. In 2014, a large-scale meta-analysis on social cognition and the cerebellum that included over 350 functional magnetic resonance imaging (fMRI) studies by Van Overwalle, Baetens, Mariën, and Vandekerckhove (2014) revealed consistent activation of the cerebellum. Cerebellar activity was present in about one-third of most social mirror and mentalizing studies, and in about all studies that involved more complex and abstract social mentalizing inferences (cf. Trope & Liberman, 2010). Abstract mentalizing involves, for instance, person trait judgments as opposed to visual descriptions of the same behaviors (e.g., respectively judging "why" versus "how" a person is reading a book; Baetens, Ma, Steen, & Van Overwalle, 2014) and inferences about the past or future as opposed to the present (Van Hoeck, Begtas, et al., 2013; Van Hoeck, Ma, et al., 2013).

The discovery of the role of the cerebellum in social thinking is in line with recent research revealing systematic neural interactions between the cerebellum and cerebral cortex (Buckner, Krienen, Castellanos, Diaz, & Yeo, 2011). These researchers found network structures in the cerebellum that are similar to the network structures of the cerebral cortex (Yeo et al., 2011). In particular, Buckner et al. (2011) clearly identified in the cerebellum distinct mentalizing and mirror cerebellar networks (as part of the larger default and somatomotor brain networks, respectively) that were directly connected to homologue networks in the cerebral cortex.

Functional connectivity between the cerebellum and cerebral cortex during social reasoning was recently confirmed in a meta-analytic connectivity study on social cognition (Van Overwalle, D'aes, & Mariën, 2015) as well as in functional and effective connectivity analyses of individual participants pooled across five fMRI studies (Van Overwalle & Mariën, 2016; Van Overwalle, Van de Steen, & Mariën, 2018). These studies revealed strong evidence for robust functional cerebro-cerebellar links during social cognition, involving the social mirror and mentalizing network. In particular, during mirror tasks, functional connectivity was found between the anterior cerebellum and two major mirror areas in the cortex (Van Overwalle et al., 2015). More importantly, during social mentalizing tasks, functional connectivity was observed between the posterior cerebellum and two major mentalizing areas in the cortex, the TPJ and mPFC (Van Overwalle, Van de Steen, & Mariën, 2018; Fig. 5). Interestingly, the connections from the cerebellum to the cortex are all negative, suggesting that they reflect some sort of error signal for the cortex. In addition, recent evidence revealed that the posterior cerebellum is also functionally connected to the social mentalizing network during autobiographical memory retrieval (Addis, Moloney, Tippett, Roberts, & Hach, 2016).

Although progress has been made in understanding the importance of the cerebellum in cognition and affect, its role in social cognition remains unexplored. To elucidate its functional role in social thinking, one theoretical perspective on the general function of the cerebellum is of particular relevance. Several authors have put forth the view that the primary function of the cerebellum is to support sequence learning and memories that underpin skilled motor acquisition, which develops slowly with practice and is inaccessible to consciousness (Ferrucci et al., 2013; Ito, 2008; Pisotta & Molinari, 2014). In this respect, the cerebellum constructs internal models of motor processes involving sequencing and planning of action in order to automate and fine-tune voluntary motor processes. These internal models are highly automatized copies from the event implications generated in the cerebral cortex that continuously sends signals to check whether an anticipated event sequence fits with current behavior and its somatosensory consequences. In this sense, the cerebellum is a "forward controller." During evolution, a more advanced function developed which allowed the cerebellum to construct internal models of pure mental processes in the form of event sequences, without involvement of overt movements and somatosensory responses (Ito, 2008; Leggio, Chiricozzi, Clausi, Tedesco, & Molinari, 2011; Pisotta & Molinari, 2014). Thus, the cerebellum regulates nonmotor mental operations in much the same way as it regulates movements (Andreasen & Pierson, 2008; Bower, 1997; Schmahmann, 1998).

Our current thinking is that this sequencing process sustained by the cerebellum crucially contributes to social cognition, by providing internal models of social sequences such as action sequences that support various representations and judgments about others. The cerebellum might play a cardinal role in learning and automatizing these action sequences in internal models that function as forward controllers to anticipate emotional and behavioral reactions from others or the self during human interaction. This mechanism likely allows humans to better anticipate action sequences and their consequences during interaction in an automatic and intuitive way and to fine-tune these anticipations, making it easier to understand behaviors and to detect violations. Consequently, this sequencing mechanism may play an important role in the organization of social events and representations.

A recent study with 11 cerebellar patients provided the first evidence that the cerebellum is crucial for the ability to understand the correct order of action sequences that require an understanding of an agent's beliefs (Van Overwalle, De Coninck, Heleven, Manto, and Mariën, 2019). In this pilot study, patients performed at typical levels on several emotion and mind attribution tasks in comparison with healthy controls, including stories involving an understanding of others' emotions or beliefs. However, they did much worse on a picture sequencing task created by Langdon and Coltheart (1999), in which participants watch cartoon-like scenarios. Each scenario is shown in four pictures like a comic-strip (Fig. 6a). They are presented in a random order, and participants have to line the pictures up in a correct chronological order. These scenarios represented mechanical, routine social script and false beliefs stories. The first two conditions reflect routine non-social and social knowledge, respectively. The last condition reflects false beliefs which, as noted earlier, involve an element of false (outdated) knowledge that is a key marker of social mentalizing. Crucially, only ordering false belief sequences revealed severe impairments among patients compared to healthy controls, while no differences were found for mechanical events or routine social scripts. Very recent fMRI findings from our lab further confirmed that this picture sequencing task recruits the mentalizing network of the cerebellum in healthy participants, and even more so for generating the correct order of action sequences that involve the understanding of others' beliefs compared to routine (non)social events (Fig. 6b; Heleven, van Dun, & Van Overwalle, 2019).

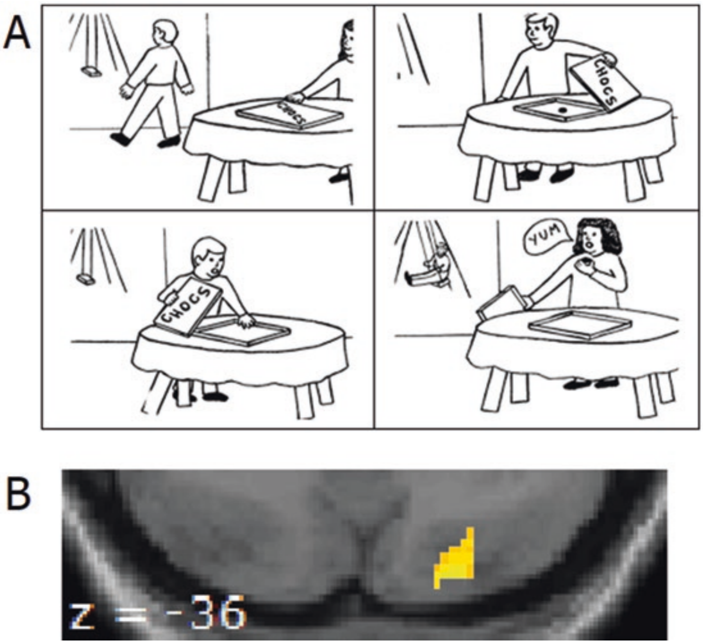


Fig. 6 Picture sequencing tasks. (**a**) An example of a false belief sequence (the correct order is 2–1–4–3). From Langdon and Coltheart (1999). (**b**) Transverse view of the posterior cerebellum showing stronger activation for false beliefs compared to social scripts and non-social (mechanical) routines (Heleven et al., 2019)

The current evidence suggests that action sequences are the backbone of cerebellar functionality in social cognition. However, this barely scratches the surface of a full insight in the cerebellum. Given its large volume, it is likely that the cerebellum contributes in many more aspects of social reasoning. Future studies can explore its role in segmenting action sequences, building knowledge in hierarchical action structures, and so on. Evidently, sequencing in action understanding is important. For instance, it makes a huge difference to learn that someone was first provoked and then became aggressive, rather than the reverse order in which someone hit first. To appreciate the importance of action sequencing, consider the idea that action sequences form the necessary cornerstone for an important capacity and evolution in humankind—story telling which united people into greater civilizations and societies rooted by shared social and religious narratives and myths that so glued together people in a united past history bound by a common faith, value, and identity (Harari, 2014).

Conclusion

In this chapter, we reviewed several lines of research on the neural underpinning of social mentalizing. First, understanding another persons' mind involves both implicit and explicit processes which originate from the same core mentalizing network, contradicting old ideas that these processes are driven by entirely different brain systems. There appears to be an implicit core process that allows observers to make quick social mentalizing inferences, presumably subserved by the TPJ and mPFC. Subsequent reprocessing allows to take in more and richer information which enable observers to verify and flexibly control their original rapid intuition, and which activate additional brain areas.

Second, one of these core mentalizing areas, the TPJ, supports here-and-now intention and belief understanding as part of an overarching joint attention function. In particular, the TPJ allows to direct one's attention to one's internal memory and extracting information from it (e.g., was the agent present or not during a critical event) while ignoring one's own perspective on external reality. Once these memorized behaviors are recalled, it is possible to make appropriate social attributions on the basic of general social knowledge.

Third, another of these mentalizing cores areas, the mPFC, encodes and represents much of this general and stabilized social knowledge on persons and traits. In effect, we see the mPFC as a hub or convergence zone of social information processing in which traits and agents represent high-level abstract summaries that integrate information on actions and behaviors at other levels of the hierarchy, through interacting with other mentalizing areas such as the TPJ.

Finally, the cerebellum contributes to many aspects of social reasoning, primarily in building internal models of action sequences that function as forward controllers to anticipate emotional and behavioral reactions from others or the self during interaction. This mechanism allows us to better anticipate action sequences and their consequences during interaction in an automatic and intuitive way and to fine-tune these anticipations, making it easier to understand behaviors and to detect violations, and to engage in social interactions.

Acknowledgments This research was funded by a SRP15 grant awarded by the Vrije Universiteit Brussel. Address for correspondence: Frank Van Overwalle, Department of Psychology, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussel, Belgium; or by e-mail: Frank.VanOverwalle@vub.ac.be.

References

Addis, D. R., Moloney, E. E. J., Tippett, L. J., Roberts, R. P., & Hach, S. (2016). Characterizing cerebellar activity during autobiographical memory retrieval: ALE and functional connectivity investigations. *Neuropsychologia, 90*, 80–93. https://doi.org/10.1016/j.neuropsychologia.2016.05.025

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*(4), 268–277. https://doi.org/10.1038/nrn1884

Andreasen, N. C., & Pierson, R. (2008). The role of the cerebellum in schizophrenia. *Biological Psychiatry, 64*(2), 81–88. https://doi.org/10.1016/j.biopsych.2008.01.003

Avidan, G., Hasson, U., Hendler, T., Zohary, E., & Malach, R. (2002). Analysis of the neuronal selectivity underlying low fMRI signals. *Current Biology: CB, 12*(12), 964–972. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12123569

Baetens, K., Ma, N., Steen, J., & Van Overwalle, F. (2014). Involvement of the mentalizing network in social and non-social high construal. *Social Cognitive and Affective Neuroscience, 9*(6), 817–824. https://doi.org/10.1093/scan/nst048

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex (New York, N.Y.: 1991), 19*(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences, 8*(12), 539–546. https://doi.org/10.1016/j.tics.2004.10.003

Bower, J. (1997). Control of sensory data acquisition. *Internatioinal Review of Neurobiology, 41*, 489–513.

Brosch, T., Schiller, D., Mojdehbakhsh, R., Uleman, J. S., & Phelps, E. A. (2013). Neural mechanisms underlying the integration of situational information into attribution outcomes. *Social Cognitive and Affective Neuroscience, 8*, 640–646. https://doi.org/10.1093/scan/nst019

Buckner, R., Krienen, F., Castellanos, A., Diaz, J. C., & Yeo, B. T. (2011). The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of Neurophysiology, 106*, 2322–2345. https://doi.org/10.1152/jn.00339.2011

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A. R., Langner, R., & Eickhoff, S. B. (2012). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure & Function, 217*(4), 783–796. https://doi.org/10.1007/s00429-012-0380-y

Cabeza, R., Ciaramelli, E., & Moscovitch, M. (2012). Cognitive contributions of the ventral parietal cortex: An integrative theoretical account. *Trends in Cognitive Sciences, 16*(6), 338–352. https://doi.org/10.1016/j.tics.2012.04.008

Chaicken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39*, 752–766.

Cunningham, W. A., & Zelazo, P. D. (2007). Attitudes and evaluations: A social cognitive neuroscience perspective. *Trends in Cognitive Sciences, 11*(3), 97–104. https://doi.org/10.1016/j.tics.2006.12.005

Cunningham, W. A., Zelazo, P. D., Packer, D. J., & Van Bavel, J. J. (2007). The iterative reprocessing model: A multilevel framework for attitudes and evaluation. *Social Cognition, 25*(5), 736–760. https://doi.org/10.1521/soco.2007.25.5.736

Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *The Neuroscientist, 13*(6), 580–593. https://doi.org/10.1177/1073858407304654

Delplanque, J., Heleven, E., & Van Overwalle, F. (2019). Neural representations of groups and stereotypes using fMRI repetition suppression. *Scientific Reports, 9*(1), 3190.

Denny, B., Ochsner, K., Weber, J., & Wager, T. D. (2014). Anticipatory brain activity predicts the success or failure of subsequent emotion regulation. *Social Cognitive and Affective Neuroscience, 9*(4), 403–411. Retrieved from http://scan.oxfordjournals.org/content/9/4/403.short

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience, 24*(8), 1742–1752. https://doi.org/10.1162/jocn_a_00233

Eickhoff, S., Laird, A., & Grefkes, C. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial. *Human Brain Mapping, 2926*, 2907–2926. https://doi.org/10.1002/hbm.20718

Epstein, S. (1994). Integration of the cognitive and psychodynamic unconscious. *American Psychologist, 49*, 709–724.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Ferrucci, R., Brunoni, A. R., Parazzini, M., Vergari, M., Rossi, E., Fumagalli, M., … Priori, A. (2013). Modulating human procedural learning by cerebellar transcranial direct current stimulation. *The Cerebellum, 12*(4), 485–492. https://doi.org/10.1007/s12311-012-0436-9

Forbes, C. E., Cox, C. L., Schmader, T., & Ryan, L. (2012). Negative stereotype activation alters interaction between neural correlates of arousal, inhibition and cognitive control. *Social Cognitive and Affective Neuroscience, 7*(7), 771–781. https://doi.org/10.1093/scan/nsr052

Forbes, C. E., & Grafman, J. (2010). The role of the human prefrontal cortex in social cognition and moral judgment. *Annual Review of Neuroscience, 33*, 299–324. https://doi.org/10.1146/annurev-neuro-060909-153230

Forbes, C. E., & Grafman, J. (2013). Social neuroscience: The second phase. *Frontiers in Human Neuroscience, 7*, 20. https://doi.org/10.3389/fnhum.2013.00020

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*(1), 21–38. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7870861

Gilead, M., Liberman, N., & Maril, A. (2013). From mind to matter: Neural correlates of abstract and concrete mindsets. *Social Cognitive and Affective Neuroscience, 9*(5), 638–645. https://doi.org/10.1093/scan/nst031

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences, 10*(1), 14–23. https://doi.org/10.1016/j.tics.2005.11.006

Harada, T., Li, Z., & Chiao, J. Y. (2010). Differential dorsal and ventral medial prefrontal representations of the implicit self modulated by individualism and collectivism: An fMRI study. *Social Neuroscience, 5*(3), 257–271. https://doi.org/10.1080/17470910903374895

Harari, U. N. (2014). *Sapiens: A brief history of humankind*. Cambridge, England: Cambridge University Press.

Heleven, E., Boukhlal, S., & Van Overwalle, F. (2018). A stranger in my brain: Neural representation for unfamiliar persons using fMRI repetition suppression. *Social Neuroscience, 13*(5), 530–540. https://doi.org/10.1080/17470919.2017.1358663

Heleven, E., van Dun, K., & Van Overwalle, F. (2019). The posterior cerebellum is involved in constructing social action sequences: An fMRI study. *Scientific Reports, 9*(1), 11110.

Heleven, E., & Van Overwalle, F. (2016). The person within: Memory codes for persons and traits using fMRI repetition suppression. *Social Cognitive and Affective Neuroscience, 11*(1), 159–171. https://doi.org/10.1093/scan/nsv100

Heleven, E., & Van Overwalle, F. (2018a) Identifying Social representations in the Brain using Repetition Suppression: A Meta-analysis.

Heleven, E., & Van Overwalle, F. (2018b). Neural representations of others in the medial prefrontal cortex do not depend on our knowledge about them. *Social Neuroscience, 14*(3), 286–299. https://doi.org/10.1080/17470919.2018.1472139

Heleven, E., & Van Overwalle, F. (2019). The neural representation of the self in relation to close others using fMRI repetition suppression. *Social Neuroscience, 14*(6), 717–728. https://doi.org/10.1080/17470919.2019.1581657

Henson, R. (2000). Neuroimaging evidence for dissociable forms of repetition priming. *Science, 287*(5456), 1269–1272. https://doi.org/10.1126/science.287.5456.1269

Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annual Review of Psychology, 60*, 653–670. https://doi.org/10.1146/annurev.psych.60.110707.163604

Ito, M. (2008). Control of mental activities by internal models in the cerebellum. *Nature Reviews Neuroscience, 9*(4), 304–313. https://doi.org/10.1038/nrn2332

Jenkins, A. C., Macrae, C. N., & Mitchell, J. P. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proceedings of the National Academy of Sciences of the United States of America, 105*(11), 4507–4512. https://doi.org/10.1073/pnas.0708785105

Kestemont, J., Vandekerckhove, M., Ma, N., Van Hoeck, N., & Van Overwalle, F. (2013). Situation and person attributions under spontaneous and intentional instructions: An fMRI study. *Social Cognitive and Affective Neuroscience, 8*(5), 481–493. https://doi.org/10.1093/scan/nss022

Keysers, C., & Gazzola, V. (2007). Integrating simulation and theory of mind: From self to social cognition. *Trends in Cognitive Sciences, 11*, 194–196.

Kovács, Á. M., Kühn, S., Gergely, G., Csibra, G., & Brass, M. (2014). Are all beliefs equal? Implicit belief attributions recruiting core brain regions of theory of mind. *PLoS One, 9*(9), e106558. https://doi.org/10.1371/journal.pone.0106558

Kovacs, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science (New York, N.Y.), 330*(6012), 1830–1834. https://doi.org/10.1126/science.1190792

Krall, S. C., Rottschy, C., Oberwelland, E., Bzdok, D., Fox, P. T., Eickhoff, S. B., … Konrad, K. (2014). The role of the right temporoparietal junction in attention and social interaction as revealed by ALE meta-analysis. *Brain Structure and Function, 220*(2), 587–604. https://doi.org/10.1007/s00429-014-0803-z

Krueger, F., Barbey, A. K., & Grafman, J. (2009). The medial prefrontal cortex mediates social event knowledge. *Trends in Cognitive Sciences, 13*(3), 103–109. https://doi.org/10.1016/j.tics.2008.12.005

Kim, H. (2017). Brain regions that show repetition suppression and enhancement: A meta-analysis of 137 neuroimaging experiments. *Human Brain Mapping, 38*(4), 1894–1913. https://doi.org/10.1002/hbm.23492

Laird, A. R., Fox, P. M., Price, C. J., Glahn, D. C., Uecker, A. M., Lancaster, J. L., … Fox, P. T. (2005). ALE meta-analysis: Controlling the false discovery rate and performing statistical contrasts. *Human Brain Mapping, 25*(1), 155–164. https://doi.org/10.1002/hbm.20136

Langdon, R., & Coltheart, M. (1999). Mentalising, schizotypy, and schizophrenia. *Cognition, 71*, 43–71.

Leggio, M. G., Chiricozzi, F. R., Clausi, S., Tedesco, A. M., & Molinari, M. (2011). The neuropsychological profile of cerebellar damage: The sequencing hypothesis. *Cortex, 47*(1), 137–144. https://doi.org/10.1016/j.cortex.2009.08.011

Letzring, T. D., Wells, S. M., & Funder, D. C. (2006). Information quantity and quality affect the realistic accuracy of personality judgment. *Journal of Personality and Social Psychology, 91*(1), 111–123. https://doi.org/10.1037/0022-3514.91.1.111

Lieberman, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology, 58*, 259–289. https://doi.org/10.1146/annurev.psych.58.110405.085654

Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., & Baron-Cohen, S. (2011). Specialization of right temporo-parietal junction for mentalizing and its relation to social impairments in autism. *NeuroImage, 56*(3), 1832–1838. https://doi.org/10.1016/j.neuroimage.2011.02.067

Ma, N., Baetens, K., Vandekerckhove, M., Kestemont, J., Fias, W., & Van Overwalle, F. (2014). Traits are represented in the medial prefrontal cortex: An fMRI adaptation study. *Social Cognitive and Affective Neuroscience, 9*(8), 1185–1192. https://doi.org/10.1093/scan/nst098

Ma, N., Vandekerckhove, M., Baetens, K., Van Overwalle, F., Seurinck, R., & Fias, W. (2012). Inconsistencies in spontaneous and intentional trait inferences. *Social Cognitive and Affective Neuroscience, 7*(8), 937–950. https://doi.org/10.1093/scan/nsr064

Ma, N., Vandekerckhove, M., Van Hoeck, N., & Van Overwalle, F. (2012). Distinct recruitment of temporo-parietal junction and medial prefrontal cortex in behavior understanding and trait identification. *Social Neuroscience, 7*(6), 591–605. https://doi.org/10.1080/17470919.2012.686925

Ma, N., Vandekerckhove, M., Van Overwalle, F., Seurinck, R., & Fias, W. (2011). Spontaneous and intentional trait inferences recruit a common mentalizing network to a different degree: Spontaneous inferences activate only its core areas. *Social Neuroscience, 6*(2), 123–138. https://doi.org/10.1080/17470919.2010.485884

Mitchell, J. P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex, 18*(2), 262–271. https://doi.org/10.1093/cercor/bhm051

Mitchell, J. P. (2009). Inferences about mental states. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 364*(1521), 1309–1316. https://doi.org/10.1098/rstb.2008.0318

Naughtin, C. K., Horne, K., Schneider, D., Venini, D., York, A., & Dux, P. E. (2017). Do implicit and explicit belief processing share neural substrates? *Human Brain Mapping, 38*(9), 4760–4772. https://doi.org/10.1002/hbm.23700

Northoff, G., Heinzel, A., Bermpohl, F., Niese, R., Pfennig, A., Pascual-Leone, A., & Schlaug, G. (2004). Reciprocal modulation and attenuation in the prefrontal cortex: An fMRI study on emotional-cognitive interaction. *Human Brain Mapping, 21*(3), 202–212. https://doi.org/10.1002/hbm.20002

Özdem, C., Brass, M., Schippers, A., Van der Cruyssen, L., & Van Overwalle, F. (2018). The neural representation of multiple mental beliefs. *Cognitive, Affective, & Behavioral Neuroscience, 19*, 1433–1443.

Özdem, C., Brass, M., Van der Cruyssen, L., & Van Overwalle, F. (2017). The overlap between false belief and spatial reorientation in the temporo-parietal junction: The role of input modality and task. *Social Neuroscience, 12*(2), 207–217. https://doi.org/10.1080/17470919.2016.1143027

Pascual-Marqui, R. D. (1999). Review of methods for solving the EEG inverse problem. *International Journal of Bioelectromagnetism, 1*, 75–86.

Pascual-Marqui, R. D., Michel, C. M., & Lehmann, D. (1994). Low resolution electromagnetic tomography: A new method for localizing electrical activity in the brain. *International Journal of Psychophysiology, 18*, 49–65.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews. Neuroscience, 8*(12), 976–987. https://doi.org/10.1038/nrn2277

Pisotta, I., & Molinari, M. (2014). Cerebellar contribution to feedforward control of locomotion. *Frontiers in Human Neuroscience, 8*, 1–5. https://doi.org/10.3389/fnhum.2014.00475

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32*(1), 3–25. https://doi.org/10.1080/00335558008248231

Qureshi, A. W., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: Evidence from a dual-task study of adults. *Cognition, 117*(2), 230–236. https://doi.org/10.1016/j.cognition.2010.08.003

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America, 98*(2), 676–682. https://doi.org/10.1073/pnas.98.2.676

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience, 18*(1), 42–55. https://doi.org/10.1038/nrn.2016.150

Rameson, L. T., Satpute, A. B., & Lieberman, M. D. (2010). The neural correlates of implicit and explicit self-relevant processing. *NeuroImage, 50*(2), 701–708. https://doi.org/10.1016/j.neuroimage.2009.12.098

Ramsey, R., & Hamilton, A. F. D. C. (2010a). Triangles have goals too: Understanding action representation in left aIPS. *Neuropsychologia, 48*(9), 2773–2776. https://doi.org/10.1016/j.neuropsychologia.2010.04.028

Ramsey, R., & Hamilton, A. F. D. C. (2010b). Understanding actors and object-goals in the human brain. *NeuroImage, 50*(3), 1142–1147. https://doi.org/10.1016/j.neuroimage.2009.12.124

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology*. New York, NY: Academic Press.

Read, S J. (1987). Constructing causal scenarios: a knowledge structure approach to causal reasoning. *Journal of Personality and Social Psychology, 52*(2), 288–302. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3559892

Read, S. J., Jones, D. K., & Miller, L. C. (1990). Traits as goal-based categories: The importance of goals in the coherence of dispositional categories. *Journal of Personality and Social Psychology, 58*(6), 1048–1061. https://doi.org/10.1037//0022-3514.58.6.1048

Reeder, G. D. (2009). Mindreading: Judgments About Intentionality and Motives in Dispositional Inference. *Psychological Inquiry, 20*(1), 1–18. https://doi.org/10.1080/10478400802615744

Reeder, G. D., Vonk, R., Ronk, M. J., Ham, J., & Lawrence, M. (2004). Dispositional attribution: multiple inferences about motive-related traits. *Journal of Personality and Social Psychology, 86*(4), 530–544. https://doi.org/10.1037/0022-3514.86.4.530

Satpute, A. B., & Lieberman, M. D. (2006). Integrating automatic and controlled processes into neurocognitive models of social cognition. *Brain Research, 1079*(1), 86–97. https://doi.org/10.1016/j.brainres.2006.01.005

Saxe, R. R., & Kanwisher, N. (2003). People thinking about thinking people—The role of the temporo-parietal junction in "theory of mind". *NeuroImage, 19*(4), 1835–1842. https://doi.org/10.1016/S1053-8119(03)00230-1

Schmahmann, J. D. (1998). Dysmetria of thought: Clinical consequences of cerebellar dysfunction on cognition and affect. *Trends in Cognitive Sciences, 2*(9), 362–371. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21227233

Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology General, 141*(3), 433–438. https://doi.org/10.1037/a0025458

Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012). Cognitive load disrupts implicit theory-of-mind processing. *Psychological Science, 23*(8), 842–847. https://doi.org/10.1177/0956797612439070

Schneider, D., Slaughter, V. P., Becker, S. I., & Dux, P. E. (2014). Implicit false-belief processing in the human brain. *NeuroImage, 101*, 268–275. https://doi.org/10.1016/j.neuroimage.2014.07.014

Schneider, W., & Shiffrin, R. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*(1), 1–66. Retrieved from http://psycnet.apa.org/journals/rev/84/1/1/

Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E. N., & Saxe, R. R. (2009). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLoS One, 4*(3), 1–7. https://doi.org/10.1371/journal.pone.0004869

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews, 42*, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology, 8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1

Tamir, D. I., & Mitchell, J. P. (2010). Neural correlates of anchoring-and-adjustment during mentalizing. *Proceedings of the National Academy of Sciences of the United States of America, 107*(24), 10827–10832. https://doi.org/10.1073/pnas.1003242107

Tamir, D. I., & Mitchell, J. P. (2012). Anchoring and adjustment during social inferences. *Journal of Experimental Psychology. General, 142*(1), 151–162. https://doi.org/10.1037/a0028232

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: A model based approach. *Social Cognitive and Affective Neuroscience, 3*(2), 119–127. https://doi.org/10.1093/scan/nsn009

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12*(12), 455–460. https://doi.org/10.1016/j.tics.2008.10.001

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*(2), 440–463. https://doi.org/10.1037/a0018963

Van der Cruyssen, L., Van Duynslaeger, M., Cortoos, A., & Van Overwalle, F. (2009). ERP time course and brain areas of spontaneous and intentional goal inferences. *Social Neuroscience, 4*(2), 165–184. https://doi.org/10.1080/17470910802253836

Van Duynslaeger, M., Van Overwalle, F., & Verstraeten, E. (2007). Electrophysiological time course and brain areas of spontaneous and intentional trait inferences. *Social Cognitive and Affective Neuroscience, 2*(3), 174–188. https://doi.org/10.1093/scan/nsm016

Van Hoeck, N., Begtas, E., Steen, J., Kestemont, J., Vandekerckhove, M., & Van Overwalle, F. (2013). False belief and counterfactual reasoning in a social environment. *NeuroImage, 90C*, 315–325. https://doi.org/10.1016/j.neuroimage.2013.12.043

Van Hoeck, N., Ma, N., Ampe, L., Baetens, K., Vandekerckhove, M., & Van Overwalle, F. (2013). Counterfactual thinking: An fMRI study on changing the past for a better future. *Social Cognitive and Affective Neuroscience, 8*(5), 556–564. https://doi.org/10.1093/scan/nss031

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30*(3), 829–858. https://doi.org/10.1002/hbm.20547

Van Overwalle, F. (2011). A dissociation between social mentalizing and general reasoning. *NeuroImage, 54*(2), 1589–1599. https://doi.org/10.1016/j.neuroimage.2010.09.043

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage, 48*(3), 564–584. https://doi.org/10.1016/j.neuroimage.2009.06.009

Van Overwalle, F., Baetens, K., Mariën, P., & Vandekerckhove, M. (2014). Social cognition and the cerebellum: A meta-analysis of over 350 fMRI studies. *NeuroImage, 86*, 554–572. https://doi.org/10.1016/j.neuroimage.2013.09.033

Van Overwalle, F., D'aes, T., & Mariën, P. (2015). Social cognition and the cerebellum: A meta-analytic connectivity analysis. *Human Brain Mapping, 36*(12), 5137–5154. https://doi.org/10.1002/hbm.23002

Van Overwalle, F., De Coninck, S., Heleven, E., Manto, M., & Mariën, P. (2019). The role of the cerebellum in reconstructing social action sequences: A pilot study. *Social, Cognitive and Affective Neuroscience, 14*(5), 549–558.

Van Overwalle, F., & Mariën, P. (2016). Functional connectivity between the cerebrum and cerebellum in social cognition: A multi-study analysis. *NeuroImage, 124*(Pt A), 248–255. https://doi.org/10.1016/j.neuroimage.2015.09.001

Van Overwalle, F., Van de Steen, F., & Mariën, P. (2018). Dynamic causal modelling of the effective connectivity between the cerebrum and cerebellum in social mentalizing across five studies. *Cognitive, Affective, & Behavioral Neuroscience, 19*(1), 211–223. https://doi.org/10.3758/s13415-018-00659-y

Van Overwalle, F., Van den Eede, S., Baetens, K., & Vandekerckhove, M. (2009). Trait inferences in goal-directed behavior: ERP timing and localization under spontaneous and intentional processing. *Social Cognitive and Affective Neuroscience, 4*(2), 177–190. https://doi.org/10.1093/scan/nsp003

Van Overwalle, F., Van Duynslaeger, M., Coomans, D., & Timmermans, B. (2011). Spontaneous goal inferences are often inferred faster than spontaneous trait inferences. *Journal of Experimental Social Psychology, 48*(1), 13–18. https://doi.org/10.1016/j.jesp.2011.06.016

Welborn, B. L., & Lieberman, M. D. (2015). Person-specific theory of mind in medial pFC. *Journal of Cognitive Neuroscience, 27*(1), 1–12. https://doi.org/10.1162/jocn_a_00700

Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology, 47*(2), 237–252. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6481615

Wood, J. N., & Grafman, J. (2003). Human prefrontal cortex: Processing and representational perspectives. *Nature Reviews Neuroscience, 4*(2), 139–147. https://doi.org/10.1038/nrn1033

Woollams, A. M. (2012). Apples are not the only fruit: The effects of concept typicality on semantic representation in the anterior temporal lobe. *Frontiers in Human Neuroscience, 6*, 1–9. https://doi.org/10.3389/fnhum.2012.00085

Yee, E., Drucker, D. M., & Thompson-Schill, S. L. (2010). fMRI-adaptation evidence of overlapping neural representations for objects related in function or manipulation. *NeuroImage, 50*(2), 753–763. https://doi.org/10.1016/j.neuroimage.2009.12.036

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., … Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology, 106*(3), 1125–1165. https://doi.org/10.1152/jn.00338.2011

The Conceptual Content of Mental Activity

Jeffrey R. Binder

This chapter discusses some phenomenological and biological links between mentalizing and general concept retrieval. As attested by this book, the neural underpinnings of our ability to hypothesize about the mental content of other intentional beings has become a topic of great interest in psychology and neuroscience. The central importance of this ability in everyday human life reflects the myriad survival advantages it conveys, which are likely reflected in somewhat specialized neurobiological representations. I argue, however, that the neural systems supporting these representations also support other types of conceptual content, placing a substantial burden of proof on any claims for functional specialization.

The ability to store and use knowledge about the world is a core feature of the human brain that has been central to our evolution and survival success, making it possible to reliably avoid known dangers and anticipate future needs by planning. People have spread across the globe and flourished through the invention of technology, including such seminal inventions as constructed shelters, farming, domestication of animals, methods for storing and preserving food, and devices for capturing and transforming energy. In each of these cases, known facts about objects and observed events were mentally manipulated, analyzed, and synthesized to create novel methods for enhancing survival. Today most adults use the same processes on a daily basis to make short- and long-term plans for beneficial future activities and solve small-scale problems. Creative analysis and synthesis of stored knowledge is used on a daily basis in the social sphere to resolve conflicts, communicate ideas, and organize groups of people.

In addition to providing a means of meeting the various exigencies of daily life, activation and manipulation of stored conceptual information provides a mechanism for such important mental activities as pleasurable recall, daydreaming,

J. R. Binder (✉)
Department of Neurology and Biophysics, Medical College of Wisconsin,
Milwaukee, WI, USA
e-mail: jbinder@mcw.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_20

reflection on art and culture, and analysis of one's own behavior and emotional responses.

Concepts and the Content of Mental Experiences

I argue that the primary contents, or "intentional objects" (Husserl, 1973/1900), of mental experience are concepts. A concept is a mental representation (which may be relatively simple or complex) resulting from generalization over many similar experiences, capturing what is common to these experiences. The concept of a concrete object like *dog*, for example, is an idealized or schematic representation of the characteristics of previously experienced dogs. Concepts like *dog* are referred to as category-level concepts because they refer to a set of unique individuals. Representations of particular individuals (e.g., *my dog Luna*), however, are also generalizations from experience and therefore concepts. Concepts have defining intrinsic features (e.g., shapes, colors, parts, movements, sounds), but also exist within a complex network of other associated concepts. The concept *dog*, for example, may have associations with concepts like *friend*, *love*, *loyalty*, *leash*, *bone*, *walk*, *breed*, *pedigree*, etc. established through co-occurrences in complex verbal and nonverbal experiences.

Concrete object concepts with verbal labels, like *dog*, have dominated much of the theoretical and empirical work on concepts (particularly in the neuroimaging world), but our vast store of concepts also includes concrete entities that are not objects (*air*, *water*, *soil*); concrete actions and events (represented in language mainly by verbs and sentences, but also by nouns like *party* and *explosion*); entities occurring as mental experiences (emotions and thoughts); quantity concepts (number, duration, and size); complex social/behavioral constructs (*honor*, *loyalty*, *democracy*, *justice*); cognitive and scientific domains (*geometry*, *law*, *philosophy*); spatial, temporal, and causal relation concepts; and many other categories. Because not all experiences are labeled with words, not all concepts have a name. The experience of satisfaction from another person's misfortune, for example, is an unnamed concept for English speakers who have not learned the word *schadenfreude*. In particular, many perceptual categories, acquired from generalization over repeated experiences, exist for which we have no names (e.g., the characteristic head shape of a particular kind of animal).

Activating a concept in the mind involves neural processing in a widely distributed brain network that represents (i.e., stores in long-term memory) and retrieves conceptual knowledge (Binder, Desai, Conant, & Graves, 2009). Since the mid-century "cognitive revolution," concept representations in the brain have been portrayed as highly abstract and localist, much like symbols in a computer program (Pylyshyn, 1984), and many authors still advocate at least a partial role for abstract representations in conceptual cognition (Dove, 2009; Mahon & Caramazza, 2008). Much behavioral and neuroimaging evidence suggests, however, that activating a concrete concept also entails activating perceptual representations of the concept in various sensory-motor modalities, such as information about its visual, tactile, auditory, or associated action features (Fernandino et al., 2016; Kiefer & Pulvermüller, 2012; Meteyard, Rodriguez Cuadrado, Bahrami, & Vigliocco, 2012). The degree to which this perceptual information becomes activated and enters awareness appears to depend on task demands. At the extreme, a visual or other sensory image may appear in awareness, but such "imagery" phenomena are best understood as a manifestation of sustained concept activation rather than a qualitatively distinct process. On this view, information stored in the brain about modality-specific (i.e., visual, auditory, tactile, action) attributes of concrete objects and events is not somehow separate from the concept representation, rather it *is* (at least a large part of) the concept representation.

The central role of concept retrieval in communication is uncontroversial: What is the purpose of communication if not to transmit concepts? If an acquaintance says, for example, "We had a good tennis game last week," it is obvious that understanding this message requires retrieval of basic knowledge about the concepts *we*, *had*, *good*, *tennis*, *game*, *last*, and *week*, and about the more specific concepts *tennis game* and *last week*. From this information, you, the hearer, might construct a mental image of the tennis game you had with the speaker, and respond by communicating labels for concepts like *I* and *agree*. Activation of conceptual knowledge, however, is not confined to the domain of verbal communication. A long tradition in linguistics and psychology linking concepts with words has obscured the fact that concept retrieval is a ubiquitous and core feature of nearly all mental activity. The paragraphs that follow discuss this point in relation to several cognitive domains usually considered to be distinct from general concept retrieval processes, all of which show considerable overlap in neuroimaging studies with both general concept retrieval networks and mentalizing networks.

Retrieval of personal episodic memories is traditionally distinguished from retrieval of concepts (semantic memory), but I argue that episodic memories are composed almost entirely of concepts. Consider that retrieval of the detailed sensory-motor events that occurred during the aforementioned tennis game, even if that were possible, would not be sufficient in itself for episodic memory retrieval. A particular set of sensory-motor events can only be recognized *as* a tennis game by retrieving the concept *tennis game*. Put another way, "understanding" always involves concept retrieval, and concepts exist in the brain to provide understanding. In the case of episodic memories, what is mainly remembered are not the detailed sensory-motor events that occurred, but an abstract version of events composed of concepts with varying amounts of perceptual detail. Episodic memory might be more properly seen as a particular kind of knowledge manipulation that creates spatial-temporal configurations of concepts representing objects, events, and other entities, including cognitive and affective phenomena.

"Autobiographical" memory is even more clearly dependent on concept retrieval, for in this case the original events have been stripped of nearly all perceptual detail and are remembered mainly as facts (e.g., place of birth, childhood home, education history). To say, "I was born in Chicago" is not to claim any perceptual memory for the events of the birth, but rather to retrieve the concepts of *birth*, *in*, and *Chicago* and to self-identify with this combination of concepts, where *I* and *self* are also nothing more or less than concepts.

The notion of autobiographical memory retrieval has relevance to the notion that some mental experiences engage a “concept of self” (Gillihan & Farah, 2005; Vogeley et al., 2001). In addition to autobiographical facts, the self-concept includes knowledge about one’s own beliefs and values, likes and dislikes, physical and cognitive characteristics, relationships to others, financial situation, personal goals, and so on. By definition, such information is of great personal relevance, and the ability to retain and retrieve such information seems to be a logical prerequisite for everyday decision-making. How could I plan my day-to-day activities without knowing my own preferences, abilities, and goals? Yet to claim self-referential processing as a special mental activity separate from concept retrieval seems difficult to justify. Are physical traits, cognitive abilities, values, relationships, and goals not concepts? To agree or disagree, for example, with the statement “I value financial independence” surely depends on the ability to retrieve a representation of the various concepts expressed in this proposition, and probably on retrieval of a wide range of associated concepts, like *parents* and *job*. As mentioned above, even the notion of “I, myself” is a concept, if an elemental one formed at a very early stage of cognitive development. The view that self-processing arises from association of the “I, myself” concept with other concepts unpacks and demystifies this seemingly special mental ability, revealing it to be yet another instance of concept retrieval and association.

Prospection, i.e., imagining the future, is often held to be a prominent component of mental experience (Ingvar, 1985; Schacter & Addis, 2007). As was the case with imagining past events (episodic memory retrieval) and reflecting on one’s concept of self, it is difficult to see how imagining future events could proceed without the core process of concept retrieval. A useful example is the participant told to “rest” in an fMRI experiment, who uses this time to consider available options for dinner after the scanning session is finished. Given our essential status as animals who benefit from the ability to store, recall, and assess food sources, it seems likely that this particular example of “future planning” has extensively evolved over the eons and provided important survival advantages. Even a cursory consideration of the processing involved, however, reveals this seemingly “special” activity of prospection to be little more than activation and evaluation of a set of related concepts. The varieties of possible cuisine, the specific shops or restaurants available and their pros and cons, the time available for a meal, the specific companions one expects to dine with and their preferences, relative differences in cost—all of these are concepts formed by generalizations from prior experiences. As with episodic memory retrieval and self-oriented cognition, imagining future scenarios cannot logically be separate from retrieval of the concepts that comprise the actual content of these mental experiences.

Working with Concepts: Selection, Analysis, and Synthesis

As outlined briefly above, mental activities generally involve the retrieval, or reactivation, of concept representations. For most such activities, however, the brain processes involved go beyond mere concept activation and include selection and manipulation of activated concepts. Selection refers to the enhanced activation, probably through an attentional mechanism, of a concept or concepts that are of greatest relevance and usefulness in a given circumstance, from among a larger set of activated concepts (Badre, Poldrack, Pare-Blagoev, Insler, & Wagner, 2005; Thompson-Schill, D'Esposito, Aguirre, & Farah, 1997). In a naming task, for example, a picture of a sheep might activate a field of concepts like *sheep*, *lamb*, *goat*, *cow*, etc., requiring selection of the most appropriate response from among these competitors. Although studied almost exclusively in the setting of overt tasks, concept selection is a basic component of all conceptual processing and likely occurs even during "spontaneous" mental activity. Consider the fMRI participant planning dinner during a "resting" interval in the scanner: concept selection occurs at every phase of this mental activity, from the focus of attention on *dinner* as opposed to other meals, to selection of *restaurants* as the search domain as opposed to other types of establishments, to the use of certain criteria and not others for assessing restaurant options, to selection of some people and not others as potential companions, and so on. This classic prospection task might be redefined (somewhat arbitrarily) as a "self-processing" task if the participant plans to dine alone and therefore focuses exclusively on self-preferences. In addition to selection of concepts like *dinner* and *restaurant*, the focus on self requires selection of "self vs. other" preferences and self-preference criteria to be given the most weight. Selection mechanisms likely also play a role during recall of personal episodic memories. Such memories are not holistic, indivisible entities, but are made up of spatiotemporal configurations of object and event concepts. The experience of such a memory typically leads to attentional focus on certain aspects of the memory and not others, i.e., selection, which determines the course of subsequent episodic recall or prospective thinking.

Analysis refers to the delineation of component features of concepts. Concepts are nearly always composed of simpler elements, such as parts of objects, distinguishable sensory features of objects, separable parts of actions, participants in events, and sequential steps within events. To solve a problem or formulate a plan, it is often necessary to decompose a retrieved concept into its component parts. Deciding which car to purchase from among many options, for example, requires analysis of the concept *car* into components like *shape*, *color*, *size*, *mileage*, *reliability*, *safety*, etc. Planning a birthday party requires analysis of the concept *birthday party* into components like *invitees*, *invitations*, *location*, *cake*, *candles*, *presents*, and so on. Each of these components is also a concept, therefore what appears intuitively to be an analytical or "breaking apart" process might be better understood as a process of activating a field of associated concepts that stand in a part-whole relationship to the parent concept. In some ways, this process is opposite to selection: whereas selection aims to focus attention on a single concept by suppressing activation of related concepts, analysis aims to activate a field of closely related concepts.

Synthesis refers to the construction of new concepts, including plans for future actions, by assembling components within schemas. A schema is a representational framework that organizes category types and relationships (Rumelhart, 1980).

Schemas are used for mental organization of complex concepts, such as events involving social interactions and spatial-temporal sequences, as well as simple object concepts. A concept like *fruit*, for example, can be represented by a schema composed of “slots” for shape, size, color, taste, juiciness, seed-type, etc. A complex concept like *party* might employ a schema with slots for location, purpose, time and duration, types of attendees, sub-events during the party and their order of occurrence, etc. We use schemas to organize and understand everyday experiences by fitting features of those experiences into pre-learned schema, sometimes leading to prejudice, confirmation bias, and other effects of stereotypical thinking (Bartlett, 1932). We use schemas to plan simple and complex behaviors, typically with slots for goal(s), actor(s), instrument(s), action(s), and patient(s) (Minsky, 1975).

Mentalizing as a Conceptual Activity

Hypothesizing about the content of other people’s thoughts and motivations is arguably a special case of the more general processes of concept retrieval, concept selection, analysis, and synthesis. From infancy we discover that we have needs that must be filled, like hunger, thirst, affection, physical comfort, sleep, and safety. We also discover various means of meeting these needs, and because these needs and means of fulfillment recur many times in many situations, they become generalized concepts that we use, consciously or not, to formulate actions. The toddler’s statement “Mommy I’m hungry” is a demonstration that the child has learned the concepts *I*, *hungry*, and *mother* and is able to select these concepts from among a field of related ones like *you*, *thirsty*, and *brother*. Analysis is demonstrated by the child’s knowledge that, along with her other characteristics, *mother* is a *giver* of *food*. Through multiple experiences in which mother (or someone else) provides the child with food, a schema develops in which the child expresses (verbally or nonverbally) a need to someone, who responds by providing something to meet the need. Synthesis occurs when the concepts *hungry* and *mother* are fit into this general schema, creating an action plan.

By the time a child is able to formulate such a plan, another critical concept will likely have been learned: the concept of having a mental plan. Concepts are generalizations learned from repeated experiences. I argue that any animal who can repeatedly form mental plans and experience the state of holding in mind a mental plan will eventually develop a concept of what it is to have a mental plan. The “experience” component is critical here. Artificial intelligence devices can be programmed to formulate action plans, though our intuition tells us that this ability alone doesn’t create in the device an “experience” of having formulated a plan. Human (and many other animal) brains are different in this critical regard: we automatically extract from complex neural activation patterns a simplified representation that can be held in short-term memory and presented to “awareness.” But this general abstraction process is the same whether the raw neural activation pattern results from an external sensory stimulus, an emotional response, or a mental event.

Learning the concept *I want* or *I believe* is not essentially different, in neurobiological terms, from learning the concept *red* or *heavy*.

Now consider what the toddler who says “Mommy I’m hungry” knows about his mother’s mental contents. It is quite likely that these contents are complex, probably including thoughts about other things she needs to do, how much food there is in the house, why her toddler is hungry so often, how much fun she had last night with her friends, etc. The toddler, on the other hand, knows only that mommy intends to get him food. How does he know this? Because he has learned from his own mental experiences the concept of having a mental plan, and he has observed on many occasions his mother executing the action of bringing food. Though not articulated overtly, the child knows (or at least expects) that his mother intends to bring him food once his own action plan (“Mommy I’m hungry”) has been executed. The fact that the child has no knowledge of the many other contents of his mother’s mind is proof that such contents must be learned through generalization over many similar experiences.

These general principles extend to all concepts acquired in the domain of social and emotional cognition. As we experience our own mental states, whether these involve desires for basic needs, emotional responses, thoughts, or simply curiosity about the environment, these recurring mental experiences evolve into generalized concepts that can be identified and articulated. Included among the core components of these concepts are our own responses and actions that result from these internal states, such as facial and body gestures that reflect emotional responses, actions taken to fulfill needs, and verbal expressions (words and phrases) that communicate the contents of our mental experience. Once conceptualized, these associated responses can be recognized in others, allowing us to infer the mental states that led to the responses, providing the basis for theory of mind. In addition to inference based on observation of others’ overt responses, we identify through experience the reliable environmental contexts that give rise to particular mental states, which then become associated with those states and can be used as additional evidence to infer mental states in others. A child’s own experiences with the emotional response caused by having a treasured toy taken away, for example, produces an association between this environmental context and the emotion of anger. A simple schema develops in which a negative emotion is experienced by sudden loss of an object. Observing another child in the same situation allows a kind of pattern completion to occur in the observer, in which the observed loss activates this previously learned schema and a representation in the observer of the likely emotional response that will occur in the other child.

The main point is that complex mental and behavioral phenomena reflecting the fact that we can infer the mental content of other intentional beings are the result of nothing more than learning through generalization over repeated similar experiences. I have elsewhere addressed the possible experiential origins of the concept of “animacy” (more properly, intentionality), which is critical for limiting the domain of possible entities to which theory of mind schema can be applied (Binder et al., 2016). We do not attribute mental states and intentions to inanimate objects, for example, because these objects do not move, show emotional responses, or communicate like intentional beings. Like our knowledge of mental states, action categories, and response schema, our knowledge of intentionality is a conceptual representation that can be activated, selected, analyzed, and synthesized with other concepts to produce action plans and inferences. Claims about processing in the domain of mentalizing and social cognition should recognize the essentially conceptual nature of these behaviors and the possibility that they are particular examples of computations (complex though they may be) arising within a more general conceptual system.

Neuroimaging Considerations

Functional neuroimaging evidence on brain systems supporting mentalizing have been expertly reviewed elsewhere (Mahy, Moses, & Pfeifer, 2014; Mar, 2011; Molenberghs, Johnson, Henry, & Mattingley, 2016; Van Overwalle, 2009) and by other contributors to this volume. Core nodes of this network include the "temporoparietal junction" (an ambiguous anatomical label usually referring to angular or supramarginal portions of the inferior parietal lobe), superior temporal sulcus, medial prefrontal cortex, lateral anterior temporal lobe, and posterior cingulate cortex. As pointed out by several authors (Andrews-Hanna, 2012; Buckner, Andrews-Hanna, & Schacter, 2008; Schilbach et al., 2012; Spreng, Mar, & Kim, 2009), this network overlaps extensively with the "default mode" network and with brain regions implicated in episodic and autobiographical memory retrieval, prospection, self-processing, and moral judgments. These latter overlaps lend support to proposals that memory retrieval, prospection, and self-processing are key components of the mental activity occurring during "resting" states (Andrews-Hanna, 2012; Buckner et al., 2008; Schacter & Addis, 2007). But what is the underlying reason for these overlaps, and why are mentalizing processes supported by virtually the same brain regions that support these other cognitive processes?

As discussed above, all of these mental activities depend on the core processes of activating stored concepts, concept selection, concept analysis, and schema-based synthesis. Sometimes ignored by social cognition researchers is a large parallel literature on single-word semantic processing showing that all of these brain regions are activated by simple contrasts like (word > matched pseudoword) and (conceptual task > matched phonologic task) (Binder et al., 2009) (Fig. 1). These contrasts, which typically use simple lexical or semantic decision tasks and neutral words drawn from a mix of conceptual categories, highlight domain-general brain areas involved in the basic processes of concept storage, retrieval and selection, analysis, and synthesis. The extensive overlap between these areas and those identified in mentalizing and other social cognition studies supports the idea that mentalizing, like most other mental activities, depends to a large extent on these domain-general conceptual processes.

A critical feature of this network is that it responds in proportion to the amount of conceptual content being processed (Binder, 2016). Activation in these areas

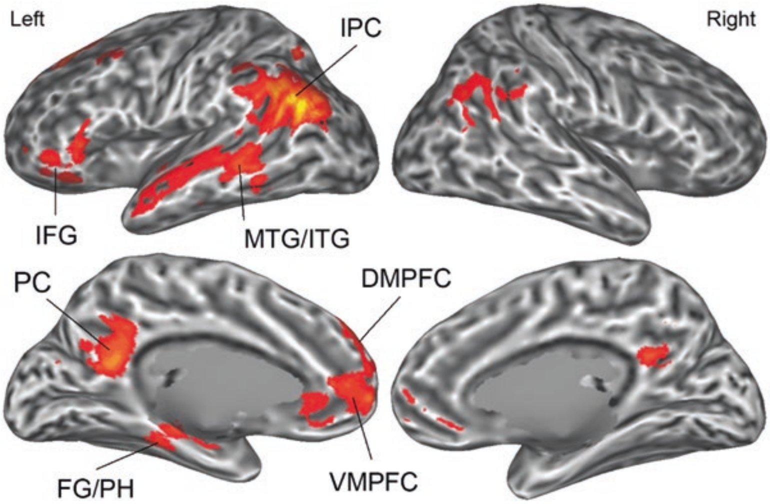


Fig. 1 A conceptual network identified by quantitative meta-analysis of 87 neuroimaging studies of semantic processing. The studies all included a manipulation of stimulus meaningfulness but no manipulation of modality-specific content. (Adapted with permission from Binder et al., 2009.) *DMPFC* dorsomedial prefrontal cortex, *FG/PH* fusiform gyrus/parahippocampus, *IFG* inferior frontal gyrus, *IPC* inferior parietal cortex, *PC* posterior cingulate/precuneus, *VMC* ventromedial prefrontal cortex

reflects the number of concepts that are active (and their intensity of activation) at any given moment, which in turn depends on the number and strength of associations that these concepts have. Distributed neural ensembles in these regions are literally equivalent to concept representations, each of which can activate a set of associated neural ensembles. All else being equal, a concept that activates many other associated concepts (causing, in turn, activation of the concepts associated with those concepts, and so on) will produce greater activation in these areas than a concept with relatively few or relatively weak associations (Bar, 2007). As mentioned above, nodes in this network are activated by single words relative to pseudowords (Binder et al., 2003; Binder, Medler, Desai, Conant, & Liebenthal, 2005; Henson, Price, Rugg, Turner, & Friston, 2002; Ischebeck et al., 2004; Kotz, Cappa, von Cramon, & Friederici, 2002; Kuchinke et al., 2005; Mechelli, Gorno-Tempini, & Price, 2003; Orfanidou, Marslen-Wilson, & Davis, 2006; Rissman, Eliassen, & Blumstein, 2003; Xiao et al., 2005). According to the present theory, this is due to the fact that pseudowords have no strong associations with concepts. Very similar results were obtained in studies comparing responses to familiar and unfamiliar proper names (Sugiura et al., 2006; Woodard et al., 2007). Like pseudowords relative to words, unfamiliar names, which refer to no known individual, have far fewer associations than familiar names, which refer to actual people about which one has associated knowledge.

Other observations explained by this general principle include activation of many of these regions by concrete relative to abstract concepts (Bedny &

Thompson-Schill, 2006; Binder et al., 2009; Binder, Medler, et al., 2005; Binder, Westbury, Possing, McKiernan, & Medler, 2005; Fliessbach, Wesi, Klaver, Elger, & Weber, 2006; Graves, Desai, Humphries, Seidenberg, & Binder, 2010; Jessen et al., 2000; Sabsevitz, Medler, Seidenberg, & Binder, 2005; Wallentin, Østergaarda, Lund, Østergaard, & Roepstorff, 2005) and frequently-used compared to infrequent words (Carreiras, Riba, Vergara, Heldmann, & Münte, 2009; Graves et al., 2010; Prabhakaran, Blumstein, Myers, Hutchison, & Britton, 2006). Concrete words show a variety of behavioral processing advantages over abstract words, including faster response times in lexical and semantic decision tasks and better recall in episodic memory tasks, reflecting the fact that concrete concepts more readily or automatically activate mental images and situational and contextual associations than abstract concepts (Paivio, 1986; Schwanenflugel, 1991). Word frequency is correlated with the number and strength of associations people generate in free association tasks (Nelson & McEvoy, 2000) and with the number of semantic features people produce in feature listing tasks (McRae, Cree, Seidenberg, & McNorgan, 2005). Assuming that words with higher frequency of use automatically activate a larger number of conceptual associations, frequency-dependent activation of the conceptual network is consistent with the aforementioned word-pseudoword, familiar-unfamiliar name, and concrete-abstract effects, all of which can be accounted for by a common underlying mechanism, i.e., relative differences in the overall intensity of activation of associated concepts.

These well-documented modulatory influences should be considered in interpreting functional imaging studies that aim to identify domain-specific processing. It is not hard to imagine, for example, the possibility that stimuli intended to specifically engage a theory of mind network might simply activate more or stronger conceptual associations than non-ToM stimuli, due to greater complexity, familiarity, or imageability, or to stronger engagement of attention by the ToM stimuli. There is no question that social interactions are an extremely important facet of our daily lives, and that we therefore know a great deal about and habitually pay close attention to human behavior. But this extended and readily accessible database of social knowledge creates an important potential confound in studies comparing processing of social vs. non-social stimuli. Are the activations observed in such comparisons specifically due to processing of social knowledge per se, or simply to stronger engagement of conceptual knowledge in general?

A concrete example of this type of confound can be found in experiments comparing verbal descriptions of complex social interactions (ToM stories) with vignettes lacking such interactions. In an item-level analysis, Dodell-Feder, Koster-Hale, Bedny, and Saxe (2011) noted substantial variation in magnitude of activation of the temporoparietal junction *within* the ToM and non-ToM conditions. That is, some ToM stimuli produced strong activation of the TPJ whereas others did not, and some non-ToM stimuli activated the region as strongly as or stronger than some ToM stimuli. As the authors noted, such variation suggests that other (non-hypothesized) stimulus features are modulating the activation. The authors considered 19 features, including 13 linguistic features (number of words per story, Flesch reading ease, anaphor reference, causal content, causal cohesion, lexical

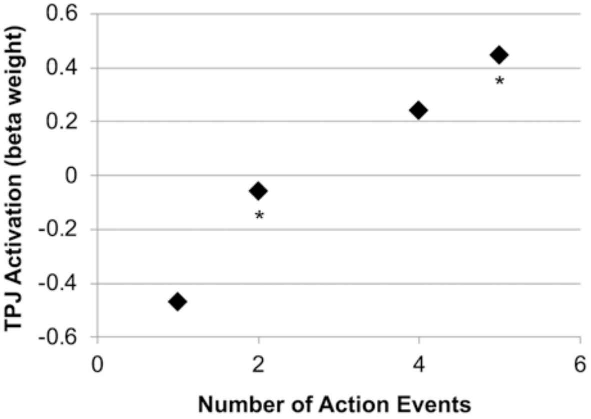


Fig. 2 Temporoparietal junction fMRI activation level produced by four story stimuli as a function of the number of action events described in each story. Items marked with an asterisk were theory-of-mind stories; unmarked items were stories describing physical events. The data are taken from examples provided in Dodell-Feder et al. (2011), Table 1

concreteness, negation, noun-phrase modification, higher-level constituency, number of words before the main verb, intentional content, attitude predication, and modality), 4 social features (number of people per story, the extent to which the items made readers think about the mental states, deception, and social status), the extent to which the items made readers think about physical causality, and the rated imageability of the events of the story. None of these features explained variation in TPJ activation. The authors did not consider the number of action events within each stimulus as a potential confound, but a cursory analysis of the four examples given in the paper suggests a relatively tight positive correlation ($r = 0.98$) between TPJ activation level and number of events portrayed (Fig. 2). Though this result needs confirmation using the entire stimulus sample, it is not unexpected given other evidence relating processing of linguistic (verbs and event nouns) and nonlinguistic markers of events with activation in the posterior temporal and inferior parietal region (Bedny, Caramazza, Grossman, Pascual-Leone, & Saxe, 2008; Bedny, Dravida, & Saxe, 2014; Zacks, Speer, Swallow, & Maley, 2010). Thus, an alternative account of some of the evidence relating mentalizing to the TPJ is that the TPJ region processes event concepts, and that ToM stimuli used in some previous studies tended to contain a higher density of event concepts compared to control stimuli.

Summary

Our ability to learn about and interact with the world through acquisition, storage, retrieval, selection, analysis, and synthesis of concept representations is a defining feature of the human brain. The intent of this chapter was to point out how these core processes underlie various mental activities that use previously acquired knowledge. My central claim is that such activities, which include language use, remembering the past, planning and envisioning the future, reflecting on the self, making moral judgments, predicting and interpreting the behavior of others, and daydreaming, are all instances in which we retrieve and manipulate concepts. It would be ridiculous, of course, to conclude somehow from this account that the study of these specific kinds of conceptual processing is not valuable and worthwhile. Understanding the specific conceptual types and relationships that support a particular domain of knowledge processing is a central goal of cognitive science. Efforts to understand the neural correlates of these processing domains, including the domain of mentalizing, would benefit from a more explicit recognition of the general conceptual processes on which they rest, tighter experimental controls, and a more cautious attitude regarding claims of functional specificity.

References

Andrews-Hanna, J. R. (2012). The brain's default network and its adaptive role in internal mentation. *The Neuroscientist, 18*(3), 251–270.

Badre, D., Poldrack, R. A., Pare-Blagoev, E. J., Insler, R. Z., & Wagner, A. D. (2005). Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex. *Neuron, 47*, 907–918.

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences, 11*(7), 280–289.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, England: Cambridge University Press.

Bedny, M., Caramazza, A., Grossman, E., Pascual-Leone, A., & Saxe, R. (2008). Concepts are more than percepts: The case of action verbs. *Journal of Neuroscience, 28*, 11347–11353.

Bedny, M., Dravida, S., & Saxe, R. (2014). Shindigs, brunches, and rodeos: The neural basis of event words. *Cognitive, Affective, & Behavioral Neuroscience, 14*(3), 891–901.

Bedny, M., & Thompson-Schill, S. L. (2006). Neuroanatomically separable effects of imageability and grammatical class during single-word comprehension. *Brain and Language, 98*, 127–139.

Binder, J. R. (2016). In defense of abstract conceptual representations. *Psychonomic Bulletin and Review, 23*, 1096–1108.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology, 33*, 130–174.

Binder, J. R., Desai, R., Conant, L. L., & Graves, W. W. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex, 19*, 2767–2796.

Binder, J. R., McKiernan, K. A., Parsons, M., Westbury, C. F., Possing, E. T., Kaufman, J. N., & Buchanan, L. (2003). Neural correlates of lexical access during visual word recognition. *Journal of Cognitive Neuroscience, 15*(3), 372–393.

Binder, J. R., Medler, D. A., Desai, R., Conant, L. L., & Liebenthal, E. (2005). Some neurophysiological constraints on models of word naming. *NeuroImage, 27*, 677–693.

Binder, J. R., Westbury, C. F., Possing, E. T., McKiernan, K. A., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience, 17*(6), 905–917.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences, 1124*, 1–38.

Carreiras, M., Riba, J., Vergara, M., Heldmann, M., & Münte, T. F. (2009). Syllable congruency and word frequency effects on brain activation. *Human Brain Mapping, 30*, 3079–3088.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. *NeuroImage, 55*(2), 705–712.

Dove, G. O. (2009). Beyond perceptual symbols: A call for representational pluralism. *Cognition, 110*, 412–431.

Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C. J., Gross, W., … Seidenberg, M. S. (2016). Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral Cortex, 26*, 2018–2034.

Fliessbach, K., Wesi, S., Klaver, P., Elger, C. E., & Weber, B. (2006). The effect of word concreteness on recognition memory. *NeuroImage, 32*, 1413–1421.

Gillihan, S. J., & Farah, M. J. (2005). Is self special? A critical review of evidence from experimental psychology and cognitive neuroscience. *Psychological Bulletin, 131*(1), 76–97.

Graves, W. W., Desai, R., Humphries, C., Seidenberg, M. S., & Binder, J. R. (2010). Neural systems for reading aloud: A multiparametric approach. *Cerebral Cortex, 20*, 1799–1815.

Henson, R. N. A., Price, C. J., Rugg, M. D., Turner, R., & Friston, K. J. (2002). Detecting latency differences in event-related BOLD responses: Application to words versus nonwords and initial versus repeated face presentations. *NeuroImage, 15*(1), 83–97.

Husserl, E. (1973/1900). *Logical investigations* (J. N. Findlay, Trans.). London, England: Routledge.

Ingvar, D. H. (1985). Memory of the future: An essay on the temporal organization of conscious awareness. *Human Neurobiology, 4*, 127–136.

Ischebeck, A., Indefrey, P., Usui, N., Nose, I., Hellwig, F., & Taira, M. (2004). Reading in a regular orthography: An fMRI study investigating the role of visual familiarity. *Journal of Cognitive Neuroscience, 16*(5), 727–741.

Jessen, F., Heun, R., Erb, M., Granath, D. O., Klose, U., Papassotiropoulos, A., & Grodd, W. (2000). The concreteness effect: Evidence for dual-coding and context availability. *Brain and Language, 74*, 103–112.

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex, 48*, 805–825.

Kotz, S. A., Cappa, S. F., von Cramon, D. Y., & Friederici, A. D. (2002). Modulation of the lexical-semantic network by auditory semantic priming: An event-related functional MRI study. *NeuroImage, 17*, 1761–1772.

Kuchinke, L., Jacobs, A. M., Grubich, C., Vo, M. L. H., Conrad, M., & Herrmann, M. (2005). Incidental effects of emotional valence in single word processing: An fMRI study. *NeuroImage, 28*, 1022–1032.

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology-Paris, 102*(1–3), 59–70.

Mahy, C. E., Moses, L. J., & Pfeifer, J. H. (2014). How and where: Theory-of-mind in the brain. *Developmental Cognitive Neuroscience, 9*, 68–81.

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology, 62*, 103–134.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature norms for a large set of living and nonliving things. *Behavior Research Methods, Instruments, & Computers, 37*(4), 547–559.

Mechelli, A., Gorno-Tempini, M. L., & Price, C. J. (2003). Neuroimaging studies of word and pseudoword reading: Consistencies, inconsistencies, and limitations. *Journal of Cognitive Neuroscience, 15*(2), 260–271.

Meteyard, L., Rodriguez Cuadrado, S., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex, 48*, 788–804.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 211–277). New York, NY: McGraw-Hill.

Molenberghs, P., Johnson, H., Henry, J. D., & Mattingley, J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience and Biobehavioral Reviews, 65*, 276–291.

Nelson, D. L., & McEvoy, C. L. (2000). What is this thing called frequency? *Memory and Cognition, 28*, 509–522.

Orfanidou, E., Marslen-Wilson, W. D., & Davis, M. H. (2006). Neural response suppression predicts repetition priming of spoken words and pseudowords. *Journal of Cognitive Neuroscience, 18*(8), 1237–1252.

Paivio, A. (1986). *Mental representations: A dual-coding approach*. New York, NY: Oxford University Press.

Prabhakaran, R., Blumstein, S. E., Myers, E. B., Hutchison, E., & Britton, B. (2006). An event-related fMRI investigation of phonological-lexical competition. *Neuropsychologia, 44*, 2209–2221.

Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. Cambridge, MA: MIT Press.

Rissman, J., Eliassen, J. C., & Blumstein, S. E. (2003). An event-related fMRI investigation of implicit semantic priming. *Journal of Cognitive Neuroscience, 15*(8), 1160–1175.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33–58). Hillsdale, NJ: Lawrence Erlbaum.

Sabsevitz, D. S., Medler, D. A., Seidenberg, M., & Binder, J. R. (2005). Modulation of the semantic system by word imageability. *NeuroImage, 27*, 188–200.

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society of London: Series B, 362*(1481), 773–786.

Schilbach, L., Bzdok, D., Timmermans, B., Fox, P. T., Laird, A. R., Vogeley, K., & Eickhoff, S. B. (2012). Introspective minds: Using ALE meta-analyses to study commonalities in the neural correlates of emotional processing, social & unconstrained cognition. *PLoS One, 7*(2), e30920–e30920.

Schwanenflugel, P. (1991). Why are abstract concepts hard to understand? In P. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 223–250). Hillsdale, NJ: Erlbaum.

Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience, 21*(3), 489–510.

Sugiura, M., Sassa, Y., Watanabe, J., Akitsuki, Y., Maeda, Y., Matsue, Y., … Kawashima, R. (2006). Cortical mechanisms of person representation: Recognition of famous and personally familiar names. *NeuroImage, 31*, 853–860.

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences USA, 94*, 14792–14797.

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30*, 829–858.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., … Zilles, K. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *NeuroImage, 14*(1 Pt 1), 170–181.

Wallentin, M., Østergaarda, S., Lund, T. E., Østergaard, L., & Roepstorff, A. (2005). Concrete spatial language: See what I mean? *Brain and Language, 92*, 221–233.

Woodard, J. L., Seidenberg, M., Nielson, K. A., Miller, S. K., Franczak, M., Antuono, P., … Rao, S. M. (2007). Temporally graded activation of neocortical regions in response to memories of different ages. *Journal of Cognitive Neuroscience, 19*(7), 1113–1124.

Xiao, Z., Zhang, J. X., Wang, X., Wu, R., Hu, X., & Tan, L. H. (2005). Differential activity in left inferior frontal gyrus for pseudowords and real words: An event-related fMRI study on auditory lexical decision. *Human Brain Mapping, 25*, 212–221.

Zacks, J. M., Speer, N. K., Swallow, K. M., & Maley, C. J. (2010). The brain's cutting-room floor: Segmentation of narrative cinema. *Frontiers in Human Neuroscience, 4*, 168.

The Role(s) of Language in Theory of Mind

Jill G. de Villiers

Introduction

What could language have to do with the development of a theory of mind? There are a number of different theoretical positions on this question. In one extreme, the answer is "nothing." That is, on that *language-as-conduit view*, theory of mind is a development in social cognition, perhaps one shared across our close primate relatives, and perhaps beginning in preverbal infants as part of core knowledge. When a child learns language, these foundational ideas become expressible. If language is delayed, a child could still pass nonverbal tests as long as the knowledge system is developed through observation of the social world added to core knowledge. If language alone is lost as in aphasia, the knowledge should remain intact. If language is tied up in adults performing dual tasks, the reasoning capacity about others' beliefs should remain.

On a second, cultural view, theory of mind is a *cultural development*, enabled by discourse that the child hears about the mind as a cause of behavior. Infants would necessarily be at only a primitive stage, perhaps capable of grasping the intention of others towards goal states, but not the contents of others' minds. Language, specifically conversation, teaches the child a theory about why people act the way they do, highlighting that discrepancies in expected behavior could result from mistakes, or ignorance. Language helps build the knowledge system, which could then, as in aphasia, survive language loss. However, language delay would imperil its development. Adults with language tied up in a dual task should be able to reason about others' belief states since the knowledge has already been built.

On a third view, an individual's language is *a cognitive tool* that assists in complex reasoning, rather like a mental scratch-pad. Learning the labels for mental states assists in this reasoning, as does the grammar, as it permits the construction

J. G. de Villiers (✉)
Psychology and Philosophy, Smith College, Northampton, MA, USA
e-mail: jdevilli@smith.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_21

of counterfactuals and conditional statements to express chains of thought that would be more awkward without language, but not impossible. For example, a succession of images might also work, or discourse of simple sentences that then chain together. Aphasic patients might be able to use some other means than language to reason, as might language-delayed children, though there should be a cost in efficiency. The development in typical children might be gradual: the greater the vocabulary and general syntactic skill, the better their theory of mind reasoning might be. It is easy to imagine that other factors such as short-term memory, and executive function, could play significant roles in this sensible, cognitive-efficiency view.

But why stop at sensible? On a fourth view, representing the opposite extreme from the conduit position, theory of mind is an outgrowth of language development. *Language provides representational structures* that scaffold belief reasoning: the semantics of intensional states arise through grammar. Infants and young children can perhaps predict goal-seeking behavior, but given syntactic development they can represent to themselves the complex propositions needed to predict and explain behavior. Language-delayed children would be impaired in this, even with tasks that make no linguistic demands. Aphasic individuals may still have sufficient access to the language faculty to so reason, but it is improbable. Adults with their language faculty occupied by another task should be unable to reason about false beliefs, as language is still needed to represent the complex propositions required for the reasoning.

Because space is limited, the focus here will be on false belief (henceforth FB) reasoning in particular, and the potential role of language in assisting that thinking in childhood. FB reasoning is sometimes considered the apex of reasoning about other minds in childhood, though even as adults we develop further nuanced understanding of other minds. In essence, FB reasoning is when the child comes to realize that an individual can believe something that is not true from the standpoint of "reality," or consensual knowledge. For example, someone might believe that they lost their glasses, when their glasses are in their hand. Or, in a classic test, a character might believe she put her chocolate in the cupboard, when we know it was subsequently moved to the refrigerator. We explain the person's searching in the cupboard by saying "she thinks her chocolate is in there." It is about the same time in life that a child can recognize that they too, might have a mistaken belief. Reviews of theory of mind developments earlier than false beliefs, and their possible links to language, can be found in de Villiers (2007).

Evidence for the Conduit View

Infants

Important evidence for the conduit view of language comes from the study of preverbal infants. Three different types of study have been done, each of them finding evidence from measures such as looking time or gaze direction that preverbal infants might be attending to another's mental states. These are called implicit theory of mind tasks, because no behavioral decision is required. In the first type, infants in the second year of life or even younger have been shown to gaze for a more protracted time at events in which a human character acts in a way contrary to expectation, in particular, a way that is not in keeping with the belief they should have formed (Onishi & Baillargeon, 2005). In particular, they are surprised when the character goes to a location where an object really is, when the character did not see it move there. In a second design, infants look expectantly at a location where a character should go, based on where that character believes something to be (Southgate, Senju, & Csibra, 2007). In a third design, very young children come to the assistance of another individual specifically if that person was not witness to how something works (Buttelman, Carpenter, & Tomasello, 2009).

The question is, are these children acting on the basis of a belief attribution, or something simpler? A number of interesting alternatives have been proposed, the most reductive of which is that the infant is responding to some accidental but correlated feature of the set-up. Another class of theories suggest that the child is responsive to a behavioral rule, such as "people go to where they last saw something" (Perner & Ruffman, 2005). However, others (Baillargeon, Buttelmann, & Southgate, 2018) have defended the sophistication of children's responses. A compromise solution by Southgate and Vernetti (2014) suggested that infants may in fact be able to follow an agent's point of view, as young as 6 months, but do not yet contrast it with their own. To put the two in contrast may only come later, perhaps when language is recruited. Finally, an important contribution comes from Apperly and Butterfill (2009)and Low and Watts (2013), who suggest that here may in fact be two systems for attending to others' theory of mind, one fast and automatic, there from very early, and independent of language or culture, and the other reflective, slow, and perhaps contingent on language acquisition. These writers claim that the behavior of infants in these tasks might have a signature limitation that distinguishes it from the success of children on explicit false belief tasks at around age four. A possible signature limitation is that infants can attend to the direction of an intention, say to a location, but are as yet unable to represent the *contents* of another's mental states, say that the object is a rock not a sponge (Low & Watts, 2013). However, at least a few experiments claim that infants can compute the contents of another's beliefs (Scott & Baillargeon, 2009).

The research on infant theory of mind, crucial to the claim that this kind of mental activity can precede language skills, is fraught with uncertainty as to the studies' replicability. There has been a general crisis of replicability in many areas of Psychology, and this particular domain suffers from special difficulty given the age of the participants and the chance that unwitting clues could be transmitted in subtle experiments. It is even more difficult when the measure might be a second or two of differential looking time, and no other behavior is available to confirm its meaning. Unfortunately there are dozens of failures reported across other labs, and as always with failures to replicate, they then do not get published (Kulke, Von Duhn, Schneider, & Rakoczy, 2018; Rakoczy, 2012; but see Baillargeon et al., 2018). This is a rich area to watch, with some brilliant innovation, but the conclusions are far from clear.

Preschool Children

If infants can succeed on implicit false belief tasks, then the conduit view faces the problem of explaining the gap that occurs between infant success and the failure of 2- and 3-year-olds on explicit false belief tasks. The answer cannot be merely that language is required to follow the task instructions or the story, because tasks have been devised (Schick, de Villiers, de Villiers, & Hoffmeister, 2007; Woolfe, Want, & Siegal, 2002) that require very little language to succeed. However, the explicit tasks do require a decision, an overt response, which differentiates them from those procedures that depend on eyegaze, an implicit response. For that reason, the common argument for the gap between infants and 4-years-olds on tasks with explicit demands has been that the younger child does not have the executive function skills to convert the implicit idea driving eyegaze into one that can mobilize a decision to act, choose the response, and resist competing demands such as reality, or a response based on the child's own beliefs. Then the real driving force of development for explicit tasks is said to be executive function skills.

There are strong findings linking executive function skills—particularly inhibitory control—to the development of false belief skills (Carlson & Moses, 2001; Hughes & Ensor, 2007). However, there are other findings in which the relationship is weak or suggest that executive function is but one skill that affects success, and when pitted against other possibilities, its contribution is not unique. In particular, it can sometimes take second place as a predictor to language skills (de Villiers et al., 2015; Farrant, Mayberry, & Fletcher, 2012; Schick et al., 2007). Furthermore, the neuroscience findings reviewed in Saxe (2009) suggest a dissociation between inhibitory control and false belief understanding in patients with brain damage.

Aphasia

Finally, powerful evidence for the conduit view comes from a small sample of aphasic patients, that is patients with significant loss of language who were tested on theory of mind. The first case reported by Varley, Siegal, and Want (2001) was of a man, SA, with significant deficits in language, who nonetheless succeeded in passing a standard verbal false belief task. Varley argued that since language and theory of mind had dissociated, it proves that in adult reasoning about beliefs, language is not recruited. However, questions have been raised about how much SA's language was lost if the directions could be followed (Baldo et al., 2005). The patient was given a verbal false belief task with reduced verbal demands. In addition, on many linguistic tasks for example, spoken word-picture matching and written word-picture matching, SA was still above chance, though grammar was impaired (Roche, 2018). Siegal and Varley (2006) reported on a second case study, MR, using a nonverbal task, and again found successful reasoning. In this case, the language tests showed impairment of the kinds of language skills normally said to be required for

ToM reasoning, for example, understanding grammar. The third and perhaps most convincing case is patient PH, studied by Apperly, Samson, Carroll, Hussain, and Humphreys (2006), who was tested on a battery of language and ToM tasks, including tests of sentential complementation. Despite impairment on the language tasks, he made virtually no errors on first and even second-order tests of FB reasoning, all nonverbal in design, but each requiring an explicit decision. Apperly et al. (2006) write: ".. a complex task, it could be presented entirely nonverbally by establishing at the outset that on every trial (including a large number of filler and control trials) PH would be asked to judge where the searcher would search." There is still an outstanding puzzle, especially for those of us who have worked with infants or profoundly language-delayed children: how were the task requirements conveyed? There is still research to be done perhaps with implicit tasks with patients with aphasia. In addition, there remains the possibility that aphasia could leave the language faculty itself intact, and affect only the performance aspects, whether production or comprehension of speech (Linebarger, Schwartz, & Saffran, 1983). If the deep aspects of language remain, they could in theory be used to explain the preserved thinking in this and other domains (see e.g., mathematics, logic: Benn, Zheng, Wilkinson, Siegal, & Varley, 2012; Fedorenko & Varley, 2016). There are differences of opinion about the nature of the "deep" aspects: Carruthers (2002) suggests that Logical Form, the logical propositions that underlie the structures in language, is the basis for human-type thinking; for Hinzen (2013), syntax is what allows human thought.

Non-humans

Perhaps the obvious place to look for evidence of theory of mind in the absence of human language is to look at non-humans. The literature is too vast to do justice to in such a chapter, but the studies suffer from many of the same difficulties as the research with infants. The work with chimpanzees and other great apes has been interpreted in a variety of ways, some generous and some, as Dennett (1983) described it, by a "killjoy" hypothesis that attributes much less intentionality to the creature's response. Tomasello and colleagues (Call & Tomasello, 2008; Krupenye, Kano, Hirata, Call, & Tomasello, 2016) have continued to support the proposition that chimpanzees can reason about others' belief states, for example, they will predict where a fellow chimpanzee (or someone dressed as one (Krupenye et al., 2016) will go to fetch a food object when that individual did not see it moved. Others, such as Povinelli and Vonk (2004), argue that empirical work suggests that chimpanzees do not even understand that "seeing leads to knowing," a precursor skill for belief reasoning. Andrews (2005) proposed that all such tests will be ambiguous, and that looking for signs that a chimpanzee seeks an *explanation* for an odd behavior might be a more fruitful approach.

Surprisingly, some tantalizing tests have been done with birds rather than apes, in particular members of the Corvid family (jays, ravens, and crows). These birds prove highly sensitive to whether another creature was watching when they hide a cache of food, and clever experimentation suggests that they are monitoring the contingent behavior of a watching conspecific (Brecht, 2017). However, in a very clever experiment, scrub-jays did not take advantage of another's false beliefs to hide food in a location that observer was led to believe was inaccessible. There is social intelligence here, but it is limited. In almost all the animal work, it is unclear how constrained these results are to hoarding food, and therefore to a specialized evolutionary path unlike that of humans.

Evidence for the Cultural View

Typical Development

We are all, at least in part, other-mind blind, because "a small but important part of the universe is enclosed within the skin of each individual" (Skinner, 1963). We have privileged access to the contents of our own minds and conscious mental states, and therefore we can only learn to describe it from others teaching us, who can only judge from our behavior. Other-mind blindness puts us at a disadvantage in teaching children words that refer to things inside of them. We can only infer that they are in the certain state say, pain as opposed to mere discomfort, happiness rather than excitement. Most writers on the subject of "private events" acknowledge that nevertheless, this is how we learn to interpret and describe the stimuli that lie inside our skins. On the cultural view**,** language about mental events has to be perceived[1] for a child learning how to express those concepts in our culture. Infants and young children have wants and feelings, and practiced caregivers can "read" their behaviors and interpret them, providing food, or assistance, or comfort. Parents usually accompany their responses with explanations and labels, saying e.g., "Oh, so you want that set of keys?" Or "Did you hurt your toe?" What this means is that it can be subject to cultural variation. Not all cultures may label behavioral reactions in the same way, or provide the same labels for the supposed internal states. As a result, we each join a discourse defined by our particular culture (Nelson, 2005). In particular, when toddlers acquire internal state language they can talk about others' feelings, preferences, desires, and perceptions. It makes new communication possible, allowing for example teasing, in addition to more positive aspects such as increased empathy (Dunn, 1988).

Through discourse, people develop a "folk psychology," that is, a lay theory about how our own minds, and then the minds of other people, operate in the world and relate to observable behavior (Hutto, 2008). We hear behavior described and explained in mental terms. For example, people talk to us about someone *trying* to get an object that they *want*, and because they *want* it, they *remember* where they

[1] I used the more neutral "perceived" because it can happen as easily in Sign as in speech.

saw it last and go to the place they last saw it. As others talk about and interpret their inner worlds, children weave these accounts into their first psychological theories (Bartsch & Wellman, 1995; Bretherton & Beeghly, 1982; Dunn & Brophy, 2005; Nelson, 2005; Shatz, Wellman, & Silber, 1983).

The cultural theory highlights the importance of hearing words as labels for underlying mental states. Dunn (e.g., Dunn & Brophy, 2005) and Meins, Fernyhough, Arnott, Leekam, and Rosnay (2013) have shown how the frequency of mental talk in a child's life influences their ability to pass false belief tasks. Some families engage in lots of mentalistic talk, and their children's reasoning is advanced; others do much less. Interestingly, one of the major contributors is family size, in particular, the presence of siblings close in age. Children from larger families have shown an advantage in theory of mind development in several studies (e.g., Astington & Jenkins, 1995; Lewis, Freeman, Kyriakidou, Maridaki-Kassotaki, & Berridge, 1996; Peterson, 2001; Ruffman, Perner, Naito, Parkin, & Clements, 1998), particularly in the area of vocabulary about emotions.

At least in Western mainstream cultures (e.g., UK, USA, Australia, Germany), an important body of research on "mind-mindedness" has revealed that the degree to which parents use language about mental states contributes to children's understanding of the mind (Dunn, Brown, Slomkowski, Tesla, & Youngblade, 1991; Meins et al., 2013; Perner, Ruffman, & Leekam, 1994; Ruffman, Slade, & Crowe, 2002). Parents also react to their child's level of understanding, for example discussing more sophisticated mental states such as "remember" increasingly often once their infants become consistent gaze followers (Slaughter, Peterson, & Carpenter, 2009). Several studies have highlighted the active role played by children in shaping their own social environments (Dunn & Plomin, 1990). How do we show that the parents are not just responding to their child's readiness to learn by speaking in a more complex way? Further research is needed to disentangle the contribution of the child's own competence, as well as genetic overlap of child and parent.

Perhaps it is not hearing just mental state terms but hearing them integrated into causally connected conversation that counts. Individual differences in preschoolers' rates of ToM development are linked to the richness of adult-child conversation involving explanations of mental states (Peterson & Slaughter, 2003; Slaughter & Peterson, 2012; Slaughter, Peterson, & Mackintosh, 2007). A recent study (Ebert, Peterson, Slaughter, & Weinert, 2017) looking at German and Australian families varying in SES replicated past studies in showing links between parents' self-reported use of elaborated mentalistic conversation and children's higher ToM scores.

Cultural and Typicality Differences

What matters most in the language input, and are there alternate routes depending on class and culture? There are some reports that propositional attitude reports are less common in the speech of some parents than others, worldwide. Several studies have found correlations between family socio-economic status (SES) and individual differences in false belief performance (e.g., Cole & Mitchell, 2000; Cutting & Dunn, 1999). Allen, de Villiers, and François (2001) investigated potential differences in white versus African American parents from different socio-economic classes in the USA, using the fairly extensive computerized transcripts from Hall, Nagy, and Linn (1984), in CHILDES (MacWhinney & Snow, 1985), of parents and their 5-year-old children. Looking just at the frequency of mental verbs, then Black working class parents produced proportionally fewer. However, Allen et al. argued if one is assessing how rich a linguistic environment is, it is important to consider more than frequency of, e.g., mental verbs; it is also necessary to consider the complexity of the contexts, and other aspects of that communicative context might compensate. For example, the Black working class children at age 5 talked much more than the other groups about communication, about who said what to whom. In doing so they used elaborate embedded language, but with no "mental" verb to be counted. Is this perhaps an important alternative route to sophisticated understanding? The use of mental terms with sentential complements is also rare among Mandarin-speaking parents and children (Snedeker & Li, 2000; Tardif & Wellman, 2000). However, Mandarin-speaking parents and children use sentential complement constructions for communication verbs (e.g., say, in Mandarin) more commonly and earlier in development than their English-speaking counterparts (Tardif & Wellman, 2000). Much more data are needed on this point cross-linguistically.

Here is yet a different perspective one can take on these findings, namely that the child is using the discourse as further evidence. Perhaps children use the language around them as a further source of evidence about other minds. In working out the meanings of words such as *think* and *know*, children may become aware of social cues that might be obscure if they just paid attention to behavior itself (e.g., Harris, 2005; Nelson, 2005). This can happen through speech or Sign. Native signing deaf children are as likely as typically developing children to engage with their parents in conversations about non-present objects, events, and ideas (Meadow, Greenberg, Erting, & Carmichael, 1981). The intact false belief comprehension shown by native signing deaf children is consistent with this richer linguistic environment (Schick et al., 2007).

A non-signing or late-signing deaf child's input might be impoverished not just because of hearing loss, but also because they have less to work with to build a theory of mind (Peterson & Siegal, 1995). But on that view, that language merely adds to the evidence available, over a longer period of time, deaf children should eventually accumulate the necessary evidence using observation of behavior. The evidence from Pyers and Senghas (2009) on adult users of the not-yet-fully-developed Nicaraguan Sign Language contradicts that assumption. These adults from the first cohort of children entering the school did not have access to more sophisticated signed input to guide them as younger children, and still did not have a rich enough language as adults, lacking mental terms in particular (Pyers, 2004). These deaf individuals in their twenties were still impaired on false belief tasks. This finding suggests that language matters for such reasoning, and plays a role above and beyond providing extra evidence for a theory about other minds.

Other Linguistic Devices

It could be argued that language is full of devices that carry perspective, such as pronouns (I, you), spatial locatives (here, there) that indicate a speaker's point of view. Other linguistic morphemes indicate the speaker's predictions about a listener's preparedness (a, the), that is, has the speaker mentioned this before to this individual? (Van Hout, Harrigan, & de Villiers, 2010). If the child surrounded by such talk from the start, why does FB reasoning take so long to learn? Yet research has failed to find strong connections between use or understanding of these devices and classic FB reasoning. Perhaps these devices offer evidence of difference in viewpoint, but not differences in truth, and it is only the latter that matter for FB reasoning (de Villiers, 2018).

In other languages, there are even more subtle devices for indicating epistemic state, such as evidentials. Evidentials grammatically mark an utterance for how the speaker knows what she is talking about: did she see it herself, or hear about it, or infer it from some clue? Research on Turkish, Bulgarian, Romani, and Tibetan as well as Korean has revealed the complex and sometimes protracted path of development of evidential morphology in children, and yet there is no compelling evidence that mastery of evidentials is linked to the onset of FB reasoning (Aksu-Ko & Alici, 2000; Aksu-Koc, Avci, Aydin, Sefer, and Yasa 2005; de Villiers & Garfield, 2017; Kyuchukov & de Villiers, 2009; Papafragou, Li, Choi, & Han, 2007).

Evidence for the Cognitive View

Typical Development

On the cognitive view, vocabulary and general grammar development *assist* theory of mind reasoning. The theoretical arguments in this domain are rather broad, and we begin with the broadest: one might argue that language skill is a proxy for verbal intelligence. That is, perhaps advanced theory of mind skills rest on the child's general intelligence, and language skills are one of the best ways to measure human intelligence. It might be argued that finding a correlation in development between language and theory of mind reflects common genetic influence on each of these cognitive domains. Against this view are the preliminary findings from Hughes and Cutting (1999) in their twin study, indicating that the genetic influence on theory of mind was largely independent from the genes involved in language ability. The language index used in this study were the verbal subtests from the Stanford Binet Intelligence Scales (Thorndike, Hagen, & Sattler, 1986), a kind of general verbal facility or verbal intelligence, not communication skills.

Alternatively, language could be considered part of the cognitive tool-kit, not so much a reflection of intelligence as an instrument of reasoning and problem-solving. The more language a child has, the more the child can use this as a tool for reasoning: perhaps even controlling impulses and thus assisting executive function, or holding things in memory, or using chains of reasoning. All of these skills would help with explicit FB reasoning, a task where a child must follow a narrative, inhibit their own knowledge, and remember the events to predict a future action.

Several studies have found that vocabulary size predicts FB performance (Happé, 1995; Milligan, Astington, & Dack, 2007). In addition, the child's general level of language measured on a standardized test has also been shown to be highly predictive of FB reasoning (Astington & Jenkins, 1999; See Milligan et al. (2007) and Farrar, Benigno, Tompkins, and Gage (2017) for meta-analyses). There is also important new evidence from a recent study by Brooks and Meltzoff (2015), who tracked the continuity in the development of children throughout the stages from gaze-following in infancy at 10.5 months to explicit FB reasoning at age 4.5 years of age. When the children were 2.5 years, their language was assessed by parental report, specifically looking at mental state vocabulary versus a matched list of non-mental state vocabulary. At the older age, children were also tested on the PPVT (a standard test of vocabulary). Controlling for their eventual verbal ability, the children's gaze-following in infancy predicted their later mental state vocabulary, but not the matched non-mental vocabulary. The parent-reported mental state terms then predicted later ToM at 4.5 years.

Delayed Language

Language appears to be a powerful mechanism for the acquisition of explicit theory of mind skills in children with autism. Several studies show that performance on theory of mind tasks in children with autism is significantly related to both lexical knowledge (Dahlgren & Trillingsgaard, 1996; Happé, 1995; Leekam & Perner, 1991; Sparrevohn & Howie, 1995) and syntactic knowledge (Tager-Flusberg, 2000; Tager-Flusberg & Joseph, 2005; Tager-Flusberg & Sullivan, 1994).

Happé has argued (Happé, 1995) that children with autism may be able to use their language skills to "hack out" a solution, rather than using the routes to FB reasoning taken by typically developing children. That is, it is suggested that the dependency of theory of mind on language might be quite different for children with autism than for other children. Happé (1995) found that the threshold of language ability sufficient for passing such tasks is much higher in children with autism than in typically developing children. Those children on the spectrum with advanced language ability may pass false belief tasks using the scaffolding that these language skills provide. Tager-Flusberg and Joseph (2005) also argued that children with autism might miss out on securely establishing the precursors of belief reasoning, being delayed on such skills as shared attention, or sensitivity to other's intentions. But those who are proficient at language might use this to scaffold their way into understanding the behavior of others.

Deaf children who are proficient native signers, especially those born to deaf parents who sign, show neither language nor theory of mind impairments, but other deaf children do (Schick et al., 2007). The consensus is that this is because of their delayed language. Peterson and Siegal (1995) found that only 50% of deaf children who were 8–13 years of age and born to hearing parents passed an unseen change-of-location task. Similarly, Russell et al. (1998) showed that non-signing deaf children who were aged 4, 9 to 16, 11 only passed a false belief task 28% of the time. In the study by Peterson, Wellman, and Liu (2005), only a third of the late-signing deaf children aged 5.5–13.2 years could pass a false belief task, but the group showed a similar ranking of five different ToM tasks to that of hearing children, albeit at a much later age. These results are echoed in other studies (Courtin, 2000; Courtin & Melot, 1998; Gale, de Villiers, de Villiers, & Pyers, 1996).

Adult Dual Task Studies

Much of the research on language and theory of mind has been developmental research, and the general assumption made by the cultural approach in particular, in all its variants, is that if language is needed at all for theory of mind, it must be just a developmental requirement. Once a theory of mind is established, then surely adults can operate without language as an intermediary. However, those who hold the cognitive view might make the case that language is a tool for such reasoning in adults as well as children.

Dual task studies have explored the possibility that language serves as a tool in reasoning even when the task is nonverbal, such as watching a brief video and predicting the ending based on a character's belief or ignorance. The methods for such a study were established by work by Hermer-Vasquez, Spelke, and Katsnelson (1999), in which they showed that adults could not reason about complex spatial arrays while shadowing a narrative. In contrast, a rhythmic shadowing task, previously calibrated against the verbal shadowing on a visual search task, did not disrupt that reasoning. Borrowing that design, Newton and de Villiers (2007) found that complex verbal shadowing but not matched rhythmic shadowing also disrupted adults' ability to reason about an agent's false beliefs. A true belief task, in which the only difference was that the character saw what happened and acted on that true belief, was not disrupted by either kind of shadowing. A follow-up study showed that shadowing non-English (Swahili) also disrupted FB reasoning, even though no meaning could have been extracted from the Swahili being shadowed (Newton, 2006).

However, the results and interpretation of adult shadowing studies continue to be mixed. Dungan and Saxe (2012) replicated the finding that adults were impaired in reasoning about beliefs while verbal shadowing, but complex rhythmic shadowing, calibrated to the skills of the individuals, also disrupted their reasoning. They therefore interpret the interference effect as due to a more general attentional disruption. Forgeot d'Arc and Ramus (2011) also found some disruption of FB reasoning for adults who were verbally shadowing, but since their participants were also affected in their *causal* reasoning, the authors rejected the possibility that the language disruption was specific to belief reasoning. Most recently, and surprisingly, Samuel, Durdevic, Legg, Lurz, and Clayton (2019) tested adults who could succeed at FB reasoning while simultaneously engaged in verbal shadowing. However, their study used verbal shadowing of simple material, an 8-digit numeric sequence. The earlier studies had used shadowing of a complex narrative from an audio book. Clearly there is more work to be done here to discover whether there is some lower limit to the complexity of the material being shadowed, in order for it to interfere. In addition, perhaps the complexity of the event matters. Events involving true beliefs prove trivial to follow (Dungan & Saxe, 2012; Newton & de Villiers, 2007), but apparently some causal events can be made as complex as those involved in FB reasoning (Forgeot d'Arc and Ramus 2011).

Would adult success on an implicit, gaze-following task be impervious to any amount of verbal interference, which seems likely given the infant results? In principle this would appear to be a simple experimental question, were it not for the substantial difficulty several laboratories have had in getting adults to gaze consistently in the expected way in infant-style implicit false belief tasks (Kulke et al., 2018; Lin, 2009). More innovation and complexity might be required to engage adult participant's attention in where balls get hidden!

A recent German study of a large number of typical adults using a complex structure equation model confirmed a significant contribution of language skills to a variety of theory of mind tasks that require reflective reasoning (Meinhardt-Injac, Daum, Meinhardt, & Persike, 2018). What is unclear is whether those skills are being recruited for that reasoning in the adults, or reflect the essential role played by language in the development of the knowledge about mind in childhood.

Evidence for the Representational View

Here evidence is reviewed that assesses whether specific syntactic achievements are necessary above and beyond the role of mental state vocabulary, rich discourse, and syntactic development in general.

Typical Development

The child's own language appears to be a key, underlying mechanism for mastery of explicit FB tasks (Astington & Baird, 2005; San Juan & Astington, 2012). The question concerns the role of complement structures, a subset of the aspects of language-as-cognitive tool that seems theoretically to have special utility in representing states of others' minds. The special property that complements (1) have relative to adjunct clauses (2), is that the embedded proposition in a verb complement can be false:

1. Arthur said that he finished the paper
2. Arthur slept after he finished the paper

Complements thus allow the expression of, e.g., mistakes and lies, or false beliefs. The complement structure is unique and only occurs under certain verbs, exclusively communication and mental state verbs. These verbs allow mention of other possible worlds in which those propositions could be true, namely, worlds in the mind of the sentence subject. Finite complements are used to express what philosophers call *propositional attitudes*.

Not only can such sentences express false propositions as belonging to another's mind or perspective, but they can also capture the particular construal of a referent that may not be known to others. For example, one person may know something under a particular description, such as "my birthday gift from my grandmother," but a friend may just know it as "the green vase." If the friend then breaks the vase, it is still true to say:

3. Your friend broke your grandmother's birthday gift to you

But it would be untrue to say:

4. Your friend thought she broke your grandmother's birthday gift to you."

Could other aspects of language play the role of complements? Specific vocabulary words exist, such as "deluded," but the word alone fails to capture the specific content of a false belief:

5. Sally was deluded
6. Sally thought the pen was a candy cane.

Discourse can perhaps do the trick, though it depends on mastery of ellipses like "so" or "that":

7. The pen was not a candy cane

Sally didn't know that

Sally didn't think so

Most intriguingly, discourse does not easily allow for recursive embedding though verb complements do:

8. The bridge was broken

Sally didn't think so

Mary knew that

Adults in English (Hollebrandse, Hobbs, de Villiers, & Roeper, 2008; Hollebrandse & Roeper, 2014) do not easily see that discourse to be equivalent in meaning to:

9. Mary knew that Sally didn't think the bridge was broken.

In sum, the special advantage of complements for capturing mental states is everywhere in the literature on propositional attitudes. But how do complements play a role in establishing the concepts of mental state?

The first studies of this aspect of language in development showed that children begin using verbs such as *think* and *know* from an early age (Diessel & Tomasello, 2001; Shatz, 1994; Shatz et al., 1983) but their first uses may be less like expressions of propositional attitudes than like stereotyped forms, often self-referent, with narrow functions:

10. I don't know (used as an escape from questioning)

11. I think it's a dog (*I think* used as *maybe*).

    Crucial for FB reasoning is the ability to describe someone else's thoughts (Bartsch & Wellman, 1995). The very first expressions of third person propositional attitudes seem to emerge around 3 or 3.5 years in spontaneous speech, and occur more rarely, e.g., in Adam's transcripts in the computerized transcripts of child language CHILDES (Brown, 1973; MacWhinney & Snow, 1985):

12. Adam: She thought that was a tiger
13. Adam: He thought I said something about window

    However, in experimental settings when children are asked to understand these forms, consistent difficulty is revealed. For instance, de Villiers (1999) arranged scenarios in which characters made statements that were either lies or mistakes, such as:

14. The woman said she found her slipper. But look, it was really a mouse.

What did the woman say she found?

Three-year-olds consistently answer "mouse," even though the answer is provided in the sentence and one can argue that no "mind reading" is necessary in the situation. Four-year-olds answer "slipper." A longitudinal study of 3–4-year-olds by de Villiers and Pyers (2002) and a very large study of children aged 4–10 years in the standardization of the DELV assessment test has exposed the time course and uniformity of this development (de Villiers, Burns, & Pearson, 2003).

It would be natural to propose that children at 3 or 4 do not yet have the conceptual resources to consider others' perspectives and mental worlds, leading to errors with false complements as a result of their failures to understand others' false beliefs. However, the reverse seems to be true. In several studies, children have been shown to understand complements before they can pass false belief tasks (de Villiers & de Villiers, 2009; de Villiers & Pyers, 2002), suggesting in the strongest claim, that such language is *prerequisite* for FB reasoning.

The finding of a strong correlation between complement mastery and FB reasoning has been documented now in several different languages: English (de Villiers & Pyers, 2002), German (Perner, Sprung, Zauner, & Haider, 2003), Danish (Knüppel, Steensgaard, & Jensen de López, 2008), and ASL (Schick et al., 2007). Aksu-Koc et al. (2005) found that production of complements predicted FB reasoning in Turkish better than evidentials did. A particularly interesting case arises with the deaf adults who learned a sign language from their peers as children in Nicaragua attending a school for the deaf established in the late 1970s. The sign language has been evolving in the hands—literally—of several generations of children over the past 40 years (Senghas, Senghas, & Pyers, 2014). Pyers (2004) asked whether the older signers, who learned a still-impoverished form of the sign as children, were able to pass nonverbal false belief tasks. By using an elicitation task, Pyers found those signers who could express propositional contents under mental state verbs were able to do FB reasoning, but those signers who did not, failed them even as adults (Pyers & Senghas, 2009).

However, there are some counter-instances to the claim that the complement mastery precedes false belief understanding. In children learning Cantonese

(Cheung, Chen, & Yeung, 2009; Cheung et al., 2004; Tardif, So, & Kaciroti, 2007), a language in which the surface markers of complementation are virtually nonexistent and there is no wh-movement, the results are less clear. Tardif et al. (2007) reported a large longitudinal study of children learning Cantonese in Hong Kong, and though she found significant correlations between complement comprehension on the de Villiers and Pyers (2002) "memory for complements" task and false belief understanding, overall the children were surprisingly poor at the complement comprehension test, even at age 6. These findings partially echoed Cheung et al. (2004). Thus the complements did not seem to be prerequisite for FB reasoning in Cantonese. One complexity worth noting is that there is a special lexical item in both Cantonese and Mandarin that means "to think falsely," and it seems as if the burden of representing false beliefs is carried more by this special lexicon than by syntax in such a language. There is much that remains to be puzzled out.

Teasing Apart Variables

In particular, is it general language (e.g., Slade & Ruffman, 2005) or a specific understanding of sentential complements (e.g., de Villiers & de Villiers, 2009) that is responsible for the breakthrough around age 4 in false belief understanding? The conclusions are ambivalent, as not all the studies included both general language and complementation measures in the same investigation. In a meta-analysis in 2007, Milligan, Astington & Dack found support for the claim that complementation was the more consistent predictor of false belief understanding, though the number of relevant studies was very limited.

Other studies since 2005 have found strong effects but may not have included both complementation and general syntax among their measures. For example, Low (2010) found that understanding sentential complements predicted standard ToM tasks in a cross-sectional sample of English-speaking children, once age, nonverbal ability and implicit false belief scores were controlled. Farrar et al. (2017) provide a useful meta-analysis of the studies to date that did compare the role of complements and general language as predictors of FB reasoning. In 10 of the 18 studies (55%) that compared both, the general language hypothesis was supported over and above the specific role of complements. These studies have used a wide variety of measures to assess "general language ability," including receptive vocabulary and different measures of syntax development. However, six of these ten studies were for Cantonese and Korean. As mentioned, mental state verbs differ in Cantonese and Mandarin compared to English, in that the distinction between true and false beliefs is carried lexically, in the verb (see Tardif et al., 2007). Nevertheless, the majority of these studies tested complements with communication verbs (except for Cheung, 2006; study 2). Thus Farrar et al. argue that even these cross-linguistic studies can be used to evaluate the relative contribution of complementation and general language.

Longitudinal studies are very rare, but they can help identify the direction of influence between the variables, as well as control for initial FB reasoning. Two early studies came to conflicting results. de Villiers and Pyers (2002) studied a small group ($N = 28$) of children over a year in preschool, and tested them at four points on a battery of theory of mind and language tests. Though they had begun the study expecting that false belief understanding might be necessary for comprehending complementation, the reverse turned out to be the case. At the time that children acquired a systematic understanding of sentential complements, then they also began to reliably pass false belief tasks.

Two larger longitudinal studies were rich enough to explore the relative contributions of vocabulary, general language, executive functioning and complements to FB reasoning in English-speaking preschool children. Farrant et al. (2012) added to the model the variable of maternal mindedness, predicting that variation in maternal input would predict children's ability on sentence complements, which would then predict false belief understanding. Their sample included 91 typically developing Australian children studied twice across a year. Importantly the effects of variation in maternal mental talk was completely mediated by the children's own competence at sentential complements, which predicted their belief ability. Cognitive flexibility was a further predictor, and the direction of effect was that sentential complement mastery predicted this executive function index rather than vice versa.

The Farrant study did not use structural equation modeling for their longitudinal portion, and had a relatively small sample size for the number of variables. We had the opportunity to test a large sample of low-income children ($N = 325$) over the course of several years as part of a preschool curricular intervention study (Lonigan et al., 2015). The children had received a large battery of language, executive function, and theory of mind measures, and these were repeated several times over the course of the study, making this an ideal group to test competing models. The results of a preliminary structural equation model looking at executive function (inhibitory control), vocabulary, and sentential complements at Time 1 and Time 2 (approximately 8 months apart) showed significant direct effects of complements, vocabulary and inhibitory control at Time 1, on FB reasoning at Time 2. In addition, there were significant indirect effects of inhibitory control and vocabulary at Time 1 on FB reasoning at Time 2, mediated through complement understanding at Time 1 (Chen, 2013; de Villiers, de Villiers, Lindley, Chen, and the School Readiness Research Consortium, 2015).

Atypical Children

If complementation is needed for ToM reasoning, then children who have not mastered them due to language delays or difficulties should struggle with ToM. We know that children with Developmental Language Disorder (DLD) display primary difficulties in formal language including complementation (Steel, Rose, & Eadie, 2016; Tuller, Henry, Sizaret, & Barthez, 2012). They are reportedly delayed in ToM, though these delays may be more subtle than those attested in ASD (Andrés-Roqueta, Adrian, Clemente, & Katsos, 2013; Holmes, 2002; Tucker, 2004). Mastery of complements by children with DLD also relates to their success at ToM (Miller, 2001). The verbal demands of the ToM tests administered in the studies are not sufficient to account for their ToM performance, as researchers have used tasks that are minimally verbal and the children still show difficulties (Nilsson & López, 2016). Complements have proven predictive of performance on minimally verbal ToM tasks for both DLD (Durrleman, Burnel, & Reboul, 2017) and ASD (Durrleman et al., 2016; Durrleman & Franck, 2015), and for deaf children with language delay (Schick et al., 2007). Farrant et al. (2012) had a sample of 31 children with language delay in their study, and the results showed that sentential complements and cognitive flexibility both predicted false beliefs in this population too.

Farrar et al. (2017) in their meta-analysis restrict attention to those studies in which both general language and complementation could be contrasted. They analyze eight studies of children with autism, deafness, or SLI, all of which indicate that language was associated with performance on false belief tasks. Complementation made an independent contribution in all of these studies except for two (e.g., Farrar et al., 2009; Lind & Bowler, 2009). In some of these populations, general language was also associated with false belief understanding. Thus, for the atypically developing children there was support for the complementation hypothesis, and Farrar et al. contend that language may be especially necessary for language-delayed children to succeed on false belief tasks.

Training Studies

The theory about sentential complements has the virtue of being falsifiable by means of experimental test, unlike many of the broader proposals. In particular, it is possible to test it via a causal intervention, namely an experimental manipulation in which one changes what children know about complementation, and see if the children improve on FB reasoning. That is, give the child the tool: does it help?

In the first such study, Hale and Tager-Flusberg (2003) took children who failed both a false belief and a sentential complement pretest and trained them in one of three conditions: direct FB reasoning, sentential complements, or relative clauses (a control group). Children trained in sentential complements were exposed only to communication verbs, allowing separation of the syntax of complementation from the lexical semantics of mental verbs. Children trained on either false belief or sentential complements significantly improved their performance on false belief tasks, whereas children trained in relative clauses showed no such improvement. What is unclear is whether the children trained on false belief directly did so without also understanding sentential complements, as the post-test arguably required them. In a second training study, Lohmann and Tomasello (2003) tested whether highlighting the nature of a deceptive object—say a candle shaped like an apple—might also improve FB reasoning. It appeared that deceptive discourse without complementation per se could suffice (Lohmann & Tomasello, 2003), though the children who received the deceptive discourse training were close to mastering complementation even on the pretest. The training condition that included training in *both* discourse about deception and sentential complements led to the most improvement in false belief understanding. Shuliang, Yanjie, and Sabbagh (2014) found that Mandarin-speaking preschool children trained on sentential complements with communication verbs showed improvement on FB reasoning, even without discourse about deception. However, they also found improvement in the conditions that used thought bubbles with representations of mistaken beliefs, despite the fact that those children did not improve on complementation. The possibility thus remains that children could succeed by other routes, as they did with direct training on false beliefs in Hale and Tager-Flusberg (2003). In sum, training complements of verbs of communication (Hale & Tager-Flusberg, 2003; Mo, Su, Sabbagh, & Jiaming, 2014) boost theory of mind reasoning in typically developing children. The optimum training may be to use complements and also deceptive objects.

The participants in these training studies were not delayed for either language or ToM, and were instead children on the cusp of developing these skills anyway. For clinical purposes, it seems important to see if enhancing complementation can boost reasoning about others' thoughts in populations where ToM and/or language is affected. This might be especially useful if atypical children show the most benefit from acquiring complements (Farrar et al., 2017). Recent work by Durrleman et al. (2019) provides the answer. In that study of French-speaking children, three groups were used, all of them chosen because they failed on pretests of both FB and sentential complementation. One group were young, typically developing children, as in the previous training studies described. A second group were children with DLD, or delayed language development, that is, cognitively typical in other respects. The third were children on the autism spectrum. The criteria for inclusion were that the participants did not yet pass complement understanding or false belief tests, though the children with DLD and autism were older, and had enough language to follow the tasks. The children were all given one of two interventions using an iPad to deliver the training and automatically score: a vocabulary training app, versus a specially designed app all that trained communication verbs with sentential complements. The results of 2 or 3 sessions per week, for 3–6 weeks of training, revealed a significant change on post-tests in both sentential complementation and on FB reasoning only in the children given the complement training, and importantly, the training was equally effective across the three participant groups, suggesting clinical usefulness. It remains to be discovered whether the gains are short term, though in this study they persisted at least until a second post-test several weeks later.

Conclusion

What does a neuroscientist need to know in this area? Theories abound about the role of language in FB reasoning, or what is standardly called theory of mind. Considering all the theories: can we distinguish them with existing data?

The cultural view that children learn from discourse about the mind can be subsumed under the representational view, in that most relevant discourse would include complements of mental states. The reverse is less clear, since children can learn from acquiring complement of communication verbs in training studies. However, no-one has proposed that the information children receive in training studies is the only input they get: surely real life is simultaneously providing information that allows them to see analogies in usage across communication and mental verbs. Yet complement syntax does seem to play a critical role, even if in Chinese languages an essential cue is carried by the head of the complement, the verb "think falsely."

The cognitive tool view is a broader version of the representational view, and there may be value in considering it as an extra perspective, e.g., the roles that language, even just labels, can play in inhibitory function or short-term memory. Without the syntax of complements, it falls short as an explanation.

The conduit view finds support in the cases of infant theory of mind and the case of aphasics. Both groups seem to be succeeding at complex theory of mind sans language. The result of each have been challenged, methodologically, so neither case is resolved. The infants cannot tell us what they think, or what drives their looking, but if it is genuinely based on reading beliefs, the result must mean that the concepts are there before language. The aphasia cases reveal nothing about ontogenesis: any of the other views could be true about development, but perhaps the reasoning about beliefs can survive language loss. However, the failure on false belief of the late- or incomplete-language learning Nicaraguan signers contradicts the conduit view.

It is likely that each theory adds something to the account, and the research area continues to be highly fruitful and innovative. More work is needed on the effects of language delay or disorder, especially with better nonverbal tasks that definitively tap reasoning about the contents of belief states. There is more work needed on the possibility of two systems, one fast and instinctive, the other reflective and guided perhaps by linguistic reasoning. Additional cross-linguistic work is needed, including languages that express mental states in less common ways, and in varieties of Sign.

References

Aksu-Ko, A., & Alici, D. M. (2000). Understanding sources of beliefs and marking of uncertainty: Child's theory of evidentiality. In E. V. Clark (Ed.), *Proceedings of the 30th Annual Child Language Conference* (pp. 123–130). Stanford, CA: Stanford University.

Aksu-Koc, A., Avci, G., Aydin, C., Sefer, N., & Yasa, Y. (2005, July). *The relation between mental verbs and ToM performance: Evidence from Turkish children*. Paper presented at IASCL, Berlin.

Allen, B.A., de Villiers, J.G. & François, S. (2001) Deficit or difference: African American children's linguistic paths towards a theory of mind. In M. Almgren, A. Barrena, M-J. Ezeizabarrena, I. Idiazabal, & B. MacWhinney (eds) Research in Child Language Acquisition: Proceedings of the 8th Conference of the International Association for the Study of Child Language. Somerville, MA: Cascadilla Press.

Andrés-Roqueta, C., Adrian, J., Clemente, R., & Katsos, N. (2013). Which are the best predictors of theory of mind delay in children with specific language impairment? *International Journal of Language & Communication Disorders, 48*(6), 726–737.

Andrews, K. (2005). Chimpanzee theory of mind: Looking in all the wrong places? *Mind & Language, 20*(5), 521–536.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953–970.

Apperly, I. A., Samson, D., Carroll, N., Hussain, S., & Humphreys, G. W. (2006). Intact first- and second-order false belief reasoning in a patient with severely impaired grammar. *Social Neuroscience, 1*(3–4), 334–348.

Astington, J., & Baird, J. (2005). *Why language matters for theory of mind*. New York, NY: Oxford University Press.

Astington, J., & Jenkins, J. (1995). Theory of mind development and social understanding. *Cognition and Emotion, 9*, 151–165.

Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Invited commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development, 46*, 112–124.

Baldo, J. V., Dronkers, N. F., Wilkins, D., Ludy, C., Raskin, P., & Kim, J. (2005). Is problem solving dependent on language? *Brain and Language, 92*(3), 240–250.

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. New York, NY: Oxford University Press.

Benn, Y., Zheng, Y., Wilkinson, I. D., Siegal, M., & Varley, R. (2012). Language in calculation: A core mechanism? *Neuropsychologia, 50*(1), 1–10.

Brecht, K. F. (2017). *A multi-facetted approach to investigating theory of mind in corvids* (Doctoral dissertation). University of Cambridge.

Bretherton, I., & Beeghly, M. (1982). Talking about internal states: The acquistion of a theory of mind. *Developmental Psychology, 18*, 906–921.

Brooks, R., & Meltzoff, A. N. (2015). Connecting the dots from infancy to childhood: A longitudinal study connecting gaze following, language, and explicit theory of mind. *Journal of Experimental Child Psychology, 130*, 67–78.

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Buttelman, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition, 112*, 337–342.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences, 12*(5), 187–192. https://doi.org/10.1016/j.tics.2008.02.010

Carlson, S., & Moses, L. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development, 72*, 1032–1053.

Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences, 25*(6), 657–674.

Chen, M. (2013). *Language, inhibitory control and false belief reasoning: A longitudinal study* (Honors thesis). Smith College.

Cheung, H., Chen, H., & Yeung, W. (2009). Relations between mental verb and false belief understanding in Cantonese-speaking children. *Journal of Experimental Child Psychology, 104*, 141–155. https://doi.org/10.1016/jeco.2009.05.004

Cheung, H., Husan-Chih, C., Creed, N., Ng, L., Wang, S., & Mo, L. (2004). Relative roles of general and complementation language in theory of mind development: Evidence from Cantonese and English. *Child Development, 75*, 1155–1170. https://doi.org/10.1111/j.1467-8624.2004.00731.x

Cole, K., & Mitchell, P. (2000). Siblings in the development of executive control and a theory of mind. *British Journal of Developmental Psychology, 18*, 279–295.

Courtin, C. (2000). The impact of sign language on the cognitive development of deaf children: The case of theories of mind. *Journal of Deaf Studies and Deaf Education, 5*, 266–276.

Courtin, C., & Melot, A. M. (1998). Development of theories of mind in deaf children. In M. Marschark & D. Clark (Eds.), *Psychological perspectives on deafness: Volume 2* (pp. 79–102). Mahwah, NJ: Lawrence Erlbaum Associates.

Cutting, A., & Dunn, J. (1999). Theory of mind, emotion understanding, language and family background: Individual differences and inter-relations. *Child Development, 70*, 853–865.

Cheung, H. (2006). False belief and language comprehension in Cantonese-speaking children. *Journal of Experimental Child Psychology, 95*(2), 79–98.

de Villiers, J., & de Villiers, P. (2009). Complements enable representation of the contents of false beliefs: The evolution of a theory of mind. In S. Foster (Ed.), *Language acquisition* (pp. 169–195). New York, NY: Palgrave McMillian.

de Villiers, J., & Garfield, J. (2017). Evidentiality, questions and the reflection principle in Tibetan: What do children learn when they learn about evidentiality? In F. N. Ketrez, A. C. Küntay, Ş. Özçalışkan, & A. Özyürek (Eds.), *Social cognition and environment in language development: Studies in honor of Ayhan Aksu-Koç. John Benjamins series: Trends in language acquisition* (pp. 113–130). Amsterdam, the Netherlands: John Benjamins.

de Villiers, J. G. (1999). On acquiring the structural representations for false complements. In B. Hollebrandse (Ed.), *New perspectives on language acquisition*. Amherst, MA: UMOP.

de Villiers, J. G. (2007). The interface of language and theory of mind. *Lingua, 117*(11), 1858–1878.

de Villiers, J. G. (2018). Perspectives on truth: The case of language and false belief reasoning. In K. Syrett & S. Arunachalam (Eds.), *Semantics in language acquisition, for the John Benjamins series trends in language acquisition* (pp. 222–245). Amsterdam, the Netherlands: John Benjamins.

de Villiers, J. G., & Pyers, J. (2002). Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development, 17*(1), 1037–1060.

de Villiers, P., de Villiers, J., Lindley, E., Chen, M., & the School Readiness Research Consortium (2015, April). *Language, inhibitory control and explicit false belief understanding: A longitudinal structural equation model*. Poster presented at SRCD.

de Villiers, P. A., Burns, F., & Pearson, B. (2003). The role of language in theory of mind development of language-impaired children: Complementing theories. In B. Beachley, A. Brown, & F. Conlin (Eds.), *Proceeding of the 27th Annual Boston University Conference on Language Development* (Vol. 1, p. 188). Somerville, MA: Cascadilla Press. https://doi.org/10.1111/j.2044-835X.2011.02072.x

Dahlgren, S., & Trillingsgaard, A. (1996). Theory of mind in non-retarded children with autism and Asperger's syndrome. A research note. *Journal of Child Psychology and Psychiatry, 37*, 463–479.

Dennett, D. C. (1983). Intentional systems in cognitive ethology: The "Panglossian paradigm" defended. *Behavioral and Brain Sciences, 6*(3), 343–390.

Diessel, H., & Tomasello, M. (2001). The acquisition of finite complement clauses in English: A usage based approach to the development of grammatical constructions. *Cognitive Linguistics, 12*, 97–141.

Dungan, J., & Saxe, R. (2012). Matched false-belief performance during verbal and nonverbal interference. *Cognitive Science, 36*(6), 1148–1156.

Dunn, J. (1988). *The beginnings of social understanding* (1st ed.). Cambridge, MA: Harvard University Press.

Dunn, J., & Brophy, M. (2005). Communication, relationships, and individual differences in children's understanding of mind. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 50–69). New York, NY: Oxford University Press.

Dunn, J., Brown, J., Slomkowski, C., Tesla, C., & Youngblade, L. (1991). Young children's understanding of other people's feelings and beliefs: Individual differences and their antecedents. *Child Development, 62*, 1352–1366.

Dunn, J., & Plomin, R. (1990). *Separate lives: Why siblings are so different*. New York, NY: Basic Books.

Durrleman, S., Burnel, M., De Villiers, J. G., Thommen, E., Yan, R., & Delage, H. (2019). The clinical impact of grammar on mentalizing: A study of children with Autism Spectrum Disorder and with Developmental Language Disorder. *Frontiers in Psychology, 10*, 2478.

Durrleman, S., Burnel, M., & Reboul, A. (2017). Theory of mind in SLI revisited: Links with syntax, comparisons with ASD. *International Journal of Language & Communication Disorders, 52*(6), 816–830.

Durrleman, S., Burnel, M., Thommen, E., Foudon, N., Sonié, S., Reboul, A., & Fourneret, P. (2016). The language-cognition interface in ASD: Complement sentences and false belief reasoning. *Research in Autism Spectrum Disorders., 21*, 109.

Durrleman, S., & Franck, J. (2015). Exploring links between language and cognition in autism spectrum disorders: Complement sentences, false belief, and executive functioning. *Journal of Communication Disorders, 54*, 15–31.

Ebert, S., Peterson, C., Slaughter, V., & Weinert, S. (2017). Links among parents' mental state language, family socioeconomic status, and preschoolers' theory of mind development. *Cognitive Development, 44*, 32–48.

Farrant, B. M., Mayberry, M. T., & Fletcher, J. (2012). Language, cognitive flexibility, and explicit false belief understanding: Longitudinal analysis in typical development and specific language impairment. *Child Development, 83*(1), 223–235. https://doi.org/10.1111/j.1467-8624.2011.01681

Farrar, J., Benigno, J., Tompkins, V., & Gage, N. (2017). Are there different pathways to explicit false belief understanding? General language and complementation in typical and atypical children. *Cognitive Development, 43*(C), 49–66.

Farrar, M. J., Johnson, B., Tompkins, V., Easters, M., Zilisi-Medus, A., & Benigno, J. P. (2009). Language and theory of mind in preschool children with specific language impairment. *Journal of Communication Disorders, 42*(6), 428–441.

Fedorenko, E., & Varley, R. (2016). Language and thought are not the same thing: Evidence from neuroimaging and neurological patients. *Annals of the New York Academy of Sciences, 1369*(1), 132.

Forgeot D'Arc, B., & Ramus, F. (2011). Belief attribution despite verbal interference. *Quarterly Journal of Experimental Psychology, 64*(5), 975–990.

Gale, E., de Villiers, P., de Villiers, J., & Pyers, J. (1996). Language and theory of mind in oral deaf children. In A. Stringfellow, D. Cahama-Amitay, E. Hughes, & A. Zukowski (Eds.), *Proceedings of the 20th Annual Boston University Conference on Language Development* (Vol. 1, pp. 213–224). Somerville, MA: Cascadilla Press.

Hale, C., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: A training study. *Developmental Science, 6*(3), 346–359.

Hall, W. S., Nagy, W. E., & Linn, R. (1984). *Spoken words: Effects of situation and social group on oral word usage and frequency*. Hillsdale, NJ: Erlbaum.

Happé, F. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development, 66*, 843–855.

Harris, P. (2005). Conversation, pretense, and theory of mind. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind*. Oxford, England: Oxford University Press.

Hermer-Vasquez, L., Spelke, E. S., & Katsnelson, A. S. (1999). Sources of flexibility in human cognition: Dual-task studies of space and language. *Cognitive Psychology, 39*, 3–36.

Hinzen, W. (2013). Narrow syntax and the language of thought. *Philosophical Psychology, 26*(1), 1–23.

Hollebrandse, B., Hobbs, K., de Villiers, J. G. & Roeper, T. (2008) Second order embedding and second order false belief. In A Gavarró & M. João Freitas (Eds.), *Language Acquisition and Development: Proceedings of GALA 2007* (pp. 270–280).

Hollebrandse, B., & Roeper, T. (2014). Empirical results and formal approaches to recursion in acquisition. In T. Roeper & M. Speas (Eds.), *Recursion: Complexity in cognition*. Dordrecht, the Netherlands: Springer.

Holmes, A. M. (2002). *Theory of mind and behavior disorders in children with specific language impairment* [Abstract]. Dissertation Abstracts International: Section B: The Sciences and Engineering, 62(11-B).

Hughes, C., & Cutting, A. (1999). Nature, nurture, and individual differences in early understanding of mind. *Psychological Science, 10*(5), 429–432.

Hughes, C., & Ensor, R. (2007). Executive function and theory of mind: Predictive relations from ages 2 to 4. *Developmental Psychology, 43*(6), 1447.

Hutto, D. (2008). *Folk psychological narratives*. Cambridge, MA: MIT/Bradford books.

Knüppel, A., Steensgaard, R., & Jensen de López, K. (2008). Mental state talk by Danish preschool children. *Nordlyd: Tromsø University Working Papers on Language & Linguistics, 34*(3), 110–130.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that others will act according to false beliefs. *Science, 354*, 110–114.

Kulke, L., Von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is implicit theory of mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science, 29*(6), 888–900.

Kyuchukov, H., & de Villiers, J. (2009). Theory of mind and evidentiality in Romani-Bulgarian bilingual children. *Psychology of Language and Communication, 13*(2), 21–34.

Leekam, S., & Perner, J. (1991). Does the autistic child have a metarepresentational deficit. *Cognition, 40*, 203–218.

Lewis, C., Freeman, N. H., Kyriakidou, C., Maridaki-Kassotaki, K., & Berridge, D. M. (1996). Social influences on false belief access: Specific sibling influences or general apprenticeship? *Child Development, 67*, 2930–2947.

Lin, Y. (2009). *Without language: Adult false belief reasoning with verbal interference*. Theses, Dissertations, and Projects. 1469. Retrieved from https://scholarworks.smith.edu/theses/1469

Lind, S. E., & Bowler, D. M. (2009). Language and theory of mind in autism spectrum disorder: The relationship between complement syntax and false belief task performance. *Journal of Autism and Developmental Disorders, 39*(6), 929–937.

Linebarger, M. C., Schwartz, M. F., & Saffran, E. M. (1983). Sensitivity to grammatical structure in so-called agrammatic aphasics. *Cognition, 13*(3), 361–392.

Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false belief understanding: A training study. *Child Development, 74*(4), 1130–1144.

Lonigan, C. J., Phillips, B. M., Clancy, J., Landry, S. H., Swank, P. R., Assel, M., … the School Readiness Consortium. (2015). Impacts of a comprehensive school readiness curriculum for preschool children at risk for educational difficulties. *Child Development, 86*(6), 1773–1793.

Low, J. (2010). Preschoolers' implicit and explicit false-belief understanding: Relations with complex syntactical mastery. *Child Development, 81*(2), 597–615.

Low, J., & Watts, J. (2013). Attributing false beliefs about object identity reveals a signature blind spot in humans' efficient mind-reading system. *Psychological Science, 24*(3), 305–311.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language, 12*(2), 271–295.

Meadow, K., Greenberg, M., Erting, C., & Carmichael, H. (1981). Interactions of deaf mothers and deaf preschool children—Comparisons with 3 other groups of deaf and hearing dyads. *American Annals of the Deaf, 126*, 454–468.

Meinhardt-Injac, B., Daum, M. M., Meinhardt, G., & Persike, M. (2018). The two-systems account of theory of mind: Testing the links to social-perceptual and cognitive abilities. *Frontiers in Human Neuroscience, 12*, 25.

Meins, E., Fernyhough, C., Arnott, B., Leekam, S., & Rosnay, M. (2013). Mind-mindedness and theory of mind: Mediating roles of language and perspectival symbolic play. *Child Development, 84*(5), 1777–1790.

Miller, C. A. (2001). False belief understanding in children with specific language impairment. *Journal of Communication Disorders, 34*, 73–86. https://doi.org/10.1016/S0021-9924(00)00042-3

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development, 78*, 622–646.

Mo, S., Su, Y., Sabbagh, M. A., & Jiaming, X. (2014). Sentential complements and false belief understanding in Chinese Mandarin-speaking preschoolers: A training study. *Cognitive Development, 29*, 50–61.

Nelson, K. (2005). Language pathways into the community of minds. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind*. Oxford, England: Oxford University Press.

Newton, A. (2006). *Verbal interference with adult false belief reasoning* (Honors thesis). Cognitive Science, Smith College.

Newton, A., & de Villiers, J. G. (2007). Thinking while talking: Adults fail non-verbal false belief reasoning. *Psychological Science, 18*(7), 574–579.

Nilsson, K., & López, K. (2016). Theory of mind in children with specific language impairment: A systematic review and meta-analysis. *Child Development, 87*(1), 143–153.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*, 255–258.

Papafragou, A., Li, P., Choi, Y., & Han, C. (2007). Evidentiality in language and cognition. *Cognition, 103*, 253–295.

Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science, 308*, 214–216.

Perner, J., Ruffman, T., & Leekam, S. R. (1994). Theory of mind is contagious: You catch it from your sibs. *Child Development, 65*(4), 1228–1238.

Perner, J., Sprung, M., Zauner, P., & Haider, H. (2003). Want that is understood well before say that, think that, and false belief: A test of de Villiers's linguistic determinism on German-speaking children. *Child Development, 74*(1), 179–188.

Peterson, C. (2001). Influence of siblings' perspectives on theory of mind. *Cognitive Development, 15*, 435–455.

Peterson, C., & Siegal, M. (1995). Deafness, conversation and theory of mind. *Journal of Child Psychology and Psychiatry, 36*, 459–474.

Peterson, C., & Slaughter, V. (2003). Opening windows into the mind: Mothers' preferences for mental state explanations and children's theory of mind. *Cognitive Development, 18*(3), 399–429.

Peterson, C. C., Wellman, H. M., & Liu, D. (2005). Steps in theory-of-mind development for children with deafness or autism. *Child Development, 76*(2), 502–517.

Povinelli, D. J., & Vonk, J. (2004). We don't need a microscope to explore the chimpanzee's mind. *Mind and Language, 19*, 1–28.

Pyers, J. (2004). *The relationship between language and false-belief understanding: Evidence from learners of an emerging sign language in Nicaragua* (Ph.D. dissertation). University of California, Berkeley, CA.

Pyers, J., & Senghas, A. (2009). Language promotes false-belief understanding: Evidence from learners of a new sign language. *Psychological Science, 20*(7), 805–812.

Rakoczy, H. (2012). Do infants have a theory of mind? *British Journal of Developmental Psychology, 30*(1), 59–74.

Roche, M. (2018). *Language and non-verbal cognition in aphasia: Insights from an eye-tracking paradigm* (Master's thesis). Universitat Pompeu-Fabra, Barcelona.

Ruffman, T., Perner, J., Naito, M., Parkin, L., & Clements, W. (1998). Older but not younger siblings facilitate false belief understanding. *Developmental Psychology, 34*, 161–174.

Ruffman, T., Slade, L., & Crowe, E. (2002). The relation between children's and mothers' mental state language and theory-of-mind understanding. *Child Development, 73*(3), 734–751.

Russell, P. A., Hosie, J. A., Gray, C. D., Scott, C., Hunter, N., Banks, J. S., & Macauley, M. C. (1998). The development of theory of mind in deaf children. *Journal of Child Psychology & Psychiatry & Allied Disciplines, 39*, 903–910.

Samuel, S., Durdevic, K., Legg, E. W., Lurz, R., & Clayton, N. S. (2019). Is language required to represent others' mental states? Evidence from beliefs and other representations. *Cognitive Science, 43*, e12710. https://doi.org/10.1111/cogs.12710

San Juan, V., & Astington, J. W. (2012). Bridging the gap between implicit and explicit understanding: How language development promotes the processing and representation of false belief. *British Journal of Developmental Psychology, 30*, 105–122. https://doi.org/10.1111/j.2044-835X.2011.02051.x

Saxe, R. (2009). Theory of mind (neural basis). In W. Banks (Ed.), *Encyclopedia of consciousness*. Cambridge, MA: MIT Press.

Schick, B., de Villiers, P., de Villiers, J., & Hoffmeister, R. (2007). Language and theory of mind: A study of deaf children. *Child Development, 78*, 376–396.

Scott, R. M., & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. *Child Development, 80*(4), 1172–1196.

Senghas, R. J., Senghas, A., & Pyers, J. E. (2014). The emergence of Nicaraguan sign language: Questions of development, acquisition, and evolution. In *Biology and knowledge revisited* (pp. 305–324). New York, NY: Routledge.

Shatz, M. (1994). Theory of mind and the development of social-linguistic intelligence in early childhood. In C. Lewis & P. Mitchell (Eds.), *Children's early understanding of mind: Origins and development*. Hilllsdale, NJ: Erlbaum.

Shatz, M., Wellman, H. M., & Silber, S. (1983). The acquisition of mental verbs: A systematic investigation of first references to mental state. *Cognition, 14*, 301–321.

Shuliang, M., Yanjie, S., & Sabbagh, M. A. (2014). Sentential complements and false belief understanding in Chinese Mandarin-speaking preschoolers: A training study. *Cognitive Development, 29*, 50–61.

Siegal, M., & Varley, R. (2006). Aphasia, language and theory of mind. *Social Neuroscience, 1*(3–4), 167–174.

Skinner, B. (1963). Behaviorism at fifty. *Science, 140*(3570), 951–958.

Slade, L., & Ruffman, T. (2005). How language does (and does not) relate to theory of mind: A longitudinal study of syntax, semantics, working memory, and false belief. *British Journal of Developmental Psychology, 23*, 141.

Slaughter, V., Peterson, C., & Mackintosh, E. (2007). Mind what mother says: Narrative input and theory of mind in typical children and those on the autism spectrum. *Child Development, 78*(3), 839–858.

Slaughter, V., & Peterson, C. C. (2012). How conversational input shapes theory of mind development in infancy and early childhood. In M. Siegal & L. Surian (Eds.), *Access to language and cognitive development* (pp. 3–22). New York, NY: Oxford University Press.

Slaughter, V., Peterson, C. C., & Carpenter, M. (2009). Maternal mental state talk and infants' early gestural communication. *Journal of Child Language, 36*, 1053–1074.

Snedeker, J., & Li, P. (2000). Can the situations in which words occur account for cross-linguistic variation in vocabulary com-position? In J. Tai, & Y. Chang (Eds.), *Proceedings of the Seventh International Symposium on Chinese Languages and Linguistics*.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by two-year-olds. *Psychological Science, 18*, 587–592.

Southgate, V., & Vernetti, A. (2014). Belief-based action prediction in preverbal infants. *Cognition, 130*(1), 1–10.

Sparrevohn, R., & Howie, P. (1995). Theory of mind in children with autistic disorder: Evidence of developmental progression and the role of verbal ability. *Journal of Child Psychology and Psychiatry, 36*, 249–263.

Steel, G., Rose, M., & Eadie, P. (2016). The production of complement clauses in children with language impairment. *Journal of Speech, Language, and Hearing Research, 59*(2), 330–341.

Tager-Flusberg, H. (2000). Language and understanding minds: Connections in autism. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds:*

Perspectives from developmental cognitive neuroscience (pp. 124–149). Oxford, England: Oxford University Press.

Tager-Flusberg, H., & Joseph, R. M. (2005). Theory of mind, language, and executive functions in autism: A longitudinal perspective. In W. Schneider, R. Schumann-Hengsteler, & B. Sodian (Eds.), *Young children's cognitive development: Interrelationships among executive functioning, working memory, verbal ability, and theory of mind* (pp. 239–257). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Tager-Flusberg, H., & Sullivan, K. (1994). A second look at second-order belief attribution in autism. *Journal of Autism and Developmental Disorders, 24*(5), 577–586.

Tardif, T., So, C. W.-C., & Kaciroti, N. (2007). Language and false belief: Evidence for general, not specific, effects in Cantonese-speaking preschoolers. *Developmental Psychology, 43*, 318–340.

Tardif, T., & Wellman, H. M. (2000). Acquisition of mental state language in Mandarin- and Cantonese-speaking children. *Developmental Psychology, 36*(1), 25–43.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford–Binet Intelligence Scales*. Chicago, IL: Riverside Publishing.

Tucker, L. (2004). *Specific language impairment and theory-of-mind: Is normal language development an essential precursor for on time theory-of-mind development?* Unpublished manuscript, University of Western Australia.

Tuller, L., Henry, C., Sizaret, E., & Barthez, M. A. (2012). Specific language impairment at adolescence: Avoiding complexity. *Applied Psycholinguistics, 33*(1), 161–184.

Van Hout, A., Harrigan, K., & de Villiers, J. G. (2010). Asymmetries in the acquisition of definite and indefinite noun phrases. In P. Hendriks & C. Koster (Eds.), *Special Issue on Asymmetries in Child Language, Lingua*, 120(8) 1973–1990.

Varley, R., Siegal, M., & Want, S. C. (2001). Severe impairment in grammar does not preclude. *Theory of Mind Neurocase, 7*(6), 489–493. https://doi.org/10.1093/neucas/7.6.489

Woolfe, T., Want, S. C., & Siegal, M. (2002). Signposts to development: Theory of mind in deaf children. *Child Development, 73*, 768–778.

Constructive Episodic Simulation: Cognitive and Neural Processes

Ruben D. I. van Genugten and Daniel L. Schacter

While daydreaming about a vacation to avoid the cold Boston winter, we might think about escaping to a beach in Mexico. Or, we can imagine taking advantage of the snow to go skiing. By mentally experiencing and testing out our possibilities before we invest resources in a specific option, we can potentially maximize benefits and minimize costs without engaging in the actual behavior (Ingvar, 1985). For example, we can imagine skiing down the mountain and hurting ourselves, then decide to avoid the potential costs associated with skiing and go to the beach instead. Once we have decided which option to pursue, imagining the situation further helps us plan for it. In this case, we can imagine the sun beaming down on the beach, then realize that we probably need to pack swimsuits and sunscreen. Simulations such as these can help us try out alternative possibilities and prepare to engage in the chosen option (Jing, Madore, & Schacter, 2017).

How is the brain able to create such simulations? To answer this question, we will begin by briefly summarizing studies using fMRI and neuropsychological approaches that have implicated a core network of brain regions that largely corresponds to the well-known *default network* (Buckner, Andrews-Hanna, & Schacter, 2008; Raichle, 2015) in the simulation of hypothetical scenarios. Next, we discuss possible reasons *why* the default network is implicated in such simulations. We then consider recent research that further informs our understanding of how this core network supports the simulation of hypothetical scenarios by using three distinct experimental techniques—*episodic specificity induction, repetition suppression*, and *transcranial magnetic stimulation*—to target specific processes that underpin future simulations. Finally, we conclude by discussing the implications of this research for our understanding of empathy and mentalizing.

R. D. I. van Genugten (✉) · D. L. Schacter
Department of Psychology, Harvard University, Cambridge, MA, USA
e-mail: ruben_vangenugten@g.harvard.edu; dls@wjh.harvard.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_22

Imagining the Future and Remembering the Past: Similarities

A large body of work shows that imagining future experiences relies on many of the same brain regions as remembering past experiences (for review, see Schacter, Addis, & Buckner, 2007; Schacter et al., 2012). For example, when participants are presented with a word cue or phrase (e.g., "beach") and are asked to imagine a specific future event or remember a past event related to the cue, many of the same regions showed similarly increased activity compared with a control task that elicits semantic and visuo-spatial processing but does not involve remembering or imagining a specific event (Addis, Wong, & Schacter, 2007). Other studies, too, have documented remarkably similar activation profiles within the default network for imagining and remembering (e.g., Addis, Pan, Vu, Laiser, & Schacter, 2009; Okuda et al., 2003; Szpunar, Watson, & McDermott, 2007). These default network regions include the medial temporal lobe, precuneus, posterior cingulate cortex, retrosplenial cortex, medial prefrontal cortex, posterior inferior parietal lobe, posterior superior temporal lobe, and lateral temporal lobe. Areas outside of the default network, such as dorsal lateral prefrontal cortex and the inferior frontal gyrus, are also active in both tasks. Together, this set of regions has been characterized as a core network that serves both remembering and imagining (Benoit & Schacter, 2015).

This extensive overlap suggests that remembering the past and imagining the future may rely on the same neural mechanisms (though differences have also been observed; for detailed review, see Schacter et al., 2012). However, the kind of overlap observed in these studies alone does not provide conclusive evidence that the same mechanism is responsible for remembering the past and imagining the future. Many tasks elicit default network activity (e.g., creativity tasks, navigation, theory of mind, memory, mind wandering, self-referential processing, and counterfactual thinking), and it is not clear that all of them involve the same neural computations (Beaty, Benedek, Silvia, & Schacter, 2016; Buckner et al., 2008; Mason et al., 2007; Ochsner et al., 2004; Schacter, Benoit, De Brigard, & Szpunar, 2015). Additional evidence is needed before concluding that remembering the past and imagining the future rely on shared processes.

Other evidence for such shared processes comes from studies of amnesic patients with medial temporal lobe damage, who have difficulty remembering specific past events. Many of these patients are also unable to imagine future and other hypothetical events to the same degree as healthy controls (e.g., Race, Keane, & Verfaellie, 2011; Tulving, 1985). For example, Hassabis, Kumaran, Vann, and Maguire (2007) asked five individuals with amnesia to vividly imagine several situations and tell the researcher everything that they imagined. When compared to controls, these individuals described imaged events that were less spatially coherent, contained fewer sensory descriptions, and had fewer items in the categories of thoughts/emotions/actions and objects/people/animals. Although these results suggest that retrieving the past and imagining hypothetical events are closely related, not all amnesic patients exhibit problems imagining future and hypothetical situations (e.g., Dede, Wixted, Hopkins, & Squire, 2016; Squire et al., 2010).

A variety of theoretical interpretations of these observations have been put forward (e.g., Buckner & Carroll, 2007; Hassabis & Maguire, 2007; Suddendorf & Corballis, 2007). Here we focus on an approach referred to as the *constructive episodic simulation hypothesis* (Schacter & Addis, 2007a, 2007b, 2020), which builds on earlier observations by Tulving (1985, 2002) implicating episodic memory in the ability to project into the future. According to this hypothesis, we construct future and other hypothetical events by flexibly retrieving and recombining elements of different episodic memories (that is, memories of specific occurrences). However, the same flexible retrieval that allows us to imagine such useful hypotheticals comes with a cost: the flexibility of memory retrieval also leads to memory distortions, such as misremembering which details come from which memories (Schacter, 2019; Schacter & Addis, 2007a, 2007b, 2020; Schacter, Guerin, & St. Jacques, 2011; for recent experimental evidence on this point, see Carpenter & Schacter, 2017, 2018).

The constructive episodic simulation hypothesis suggests that episodic memory retrieval plays an important role in various forms of cognition that rely on imagining specific situations. For example, episodic retrieval is hypothesized to contribute to planning steps to achieve a personal goal (autobiographical planning; e.g., Spreng, Stevens, Chamberlain, Gilmore, & Schacter, 2010), estimating one's response to a future event (affective forecasting; Gilbert & Wilson, 2007), and imagining alternatives to a specific past personal event (episodic counterfactual thinking; e.g., De Brigard, Addis, Ford, Schacter, & Giovanello, 2013). Consistent with this view, autobiographical planning engages the default network (Gerlach, Spreng, Gilmore, & Schacter, 2011; Spreng et al., 2010; Spreng, Gerlach, Turner, & Schacter, 2015). Likewise, episodic counterfactual thinking elicits activity in many of the same brain regions as recalling the past does (De Brigard et al., 2013; Schacter et al., 2015). Such studies provide evidence consistent with the idea that episodic memory retrieval contributes to different forms of imagination. However, as noted earlier, default network activity is elicited by many different processes, so we must be careful to infer from default network activity that episodic memory contributes to these forms of imagination (Poldrack, 2006). Research discussed in the following sections of this chapter provides stronger evidence that episodic retrieval contributes to future imagining and related kinds of mental simulations. We will consider, in turn, studies that have relied on the techniques of the episodic specificity induction (ESI), repetition suppression (RS), and transcranial magnetic stimulation (TMS).

Episodic Specificity Induction: Identifying Contributions of Episodic Retrieval to Imagination

The constructive episodic simulation hypothesis states that we imagine future events in part by retrieving episodic details. To test this hypothesis, Madore, Gaesser, and Schacter (2014) developed a manipulation to temporarily boost episodic retrieval. If imagining specific future events draws on episodic retrieval, the manipulation (when compared to the control manipulation) should enhance task performance. The procedure that was developed, known as the episodic specificity induction (ESI), has proven useful for identifying episodic retrieval contributions to a variety of tasks.

The ESI is adapted from the cognitive interview, which was designed to elicit detailed memories from eyewitnesses (Fisher & Geiselman, 1992). In the ESI procedure, participants are given a brief training in retrieving episodic details from a recent event. Participants first watch a brief video and are then asked to retrieve information about the surroundings and objects in the video, the appearance of individuals, and all the actions in chronological order. Following this procedure, participants perform the task of interest (e.g., imagining future events). The effect of this ESI on the subsequent task is then compared to the effect of a control induction, which in most experiments consists of an interview about the participant's general impressions of the video (for full interview scripts, see Madore et al., 2014).

The critical need for the ESI procedure is illustrated by earlier work on the relationship between episodic memory and imagination. For example, several experiments had indicated that older adults, who provide fewer episodic details than young adults when remembering past experiences, also provide fewer episodic details when imagining future experiences (Addis, Musicaro, Pan, & Schacter, 2010; Addis, Wong, & Schacter, 2008). However, a subsequent study showed that when asked to describe a picture—a task that should not involve episodic retrieval—older adults generated fewer details that were physically present in the picture than younger adults (Gaesser, Sacchetti, Addis, & Schacter, 2011). These findings suggest that the link between remembering past experiences and imagining future experiences could be at least partially explained by factors other than episodic retrieval, such as the manner in which people talk about their experiences in the present, past, or future. Studies that do not take account of such non-episodic influences are therefore inadequate for assessing the contributions of episodic retrieval to such cognitive tasks as future imagining because these non-episodic influences may also contribute to task performance. The ESI overcomes these limitations by manipulating episodic retrieval, thereby allowing researchers to assess the downstream impact of this manipulation on subsequent tasks.

Episodic Retrieval Contributes to Future Imagining: Support for the Constructive Episodic Simulation Hypothesis

In the first study to develop and use the ESI (Madore et al., 2014), young and old adults were asked to imagine future events, remember past events, and to describe pictures. Madore et al. predicted that the two tasks hypothesized to rely on episodic retrieval—remembering the past and imagining the future—would benefit from the ESI (when compared to the control induction), while there would be no effect of the ESI on the non-episodic picture description task. Details on all three tasks were coded using procedures from the well-established autobiographical interview (Levine, Svoboda, Hay, Winocur, & Moscovitch, 2002), which distinguishes between two types of details that people provide on autobiographical tasks: internal or episodic details (e.g., who, what, where, when) and external details (e.g., semantic details, off-task comments, and repetitive details). For the picture description task, internal details were defined as details physically present in the picture, and external details were the same as in the other tasks. Madore et al. (2014) predicted and found an interaction between induction type, detail type, and task: internal/episodic details were selectively increased by the ESI for both young and old adults relative to the control induction when participants remembered past experiences and imagined future experiences, but not when they described pictures, and the number of external details did not differ between the two inductions on any of the three tasks. This pattern of results provides evidence that episodic retrieval contributes to remembering the past and imagining the future and is inconsistent with the hypothesis that ESI simply changes narrative style or the amount that participants talked. These findings are further bolstered by a subsequent experiment that yielded identical patterns of results using words rather than pictures to cue memory and imagination, and a non-episodic control task that required generating sentences and definitions in response to word cues (Madore & Schacter, 2016). Taken together, these studies support the conclusion that the ESI serves as a tool to selectively manipulate the contributions of episodic retrieval to a cognitive task such as future imagining, which is not normally considered an "episodic memory task."

Using the ESI to Identify Contributions of Episodic Retrieval to Problem Solving and Divergent Thinking

Recent studies have suggested that other tasks involving mental simulation that would not ordinarily be considered "episodic memory tasks" nonetheless draw on episodic memory retrieval. For example, Sheldon, McAndrews, and Moscovitch (2011) showed that populations with impaired episodic memory provide fewer relevant steps to solve open-ended social problems. To provide an additional test of whether episodic retrieval contributes to this kind of problem solving, Madore and Schacter (2014) administered the ESI and a control induction to young and old participants before they engaged in the means-end problem solving task used previously by Sheldon et al. (2011). During this problem solving task, which was based on earlier work by Platt and Spivack (1975), participants were asked to produce a series of problem solving steps in response to cues such as "J is having trouble getting along with the boss on his job. J is very unhappy about this. The story ends with J's boss liking him. You begin the story where J isn't getting along with his boss."

Both young and old participants produced more relevant steps to solve the problem after the ESI than the control induction, while there was no effect of the ESI on generating irrelevant steps. These results provide strong evidence that episodic retrieval contributes to means-end problem solving. Subsequent work shows that the ability to generate more steps to solve a problem has further impacts on mental well-being. After an ESI, young adults were able to imagine more steps to solve a personally worrisome problem than after a control induction, and they also reported feeling less anxious about the event (Jing, Madore, & Schacter, 2016).

Other tasks that require the generation of specific mental scenarios may likewise benefit from episodic retrieval. For instance, there is suggestive evidence that divergent thinking, or the ability to combine old elements to generate creative new ideas, may be linked with episodic retrieval. Duff, Kurczek, Rubin, Cohen, and Tranel (2013) reported that hippocampal amnesic patients show decreased performance on a battery of divergent thinking tasks when compared to controls. In addition, individual differences in divergent thinking are correlated with differences in episodic detail generation for future events (though not past events; Addis, Pan, Musicaro, & Schacter, 2016). To provide even stronger evidence for a link between episodic retrieval and divergent thinking, recent studies have administered the ESI procedure prior to divergent thinking tests. In one study (Madore, Addis, & Schacter, 2015), participants were asked to generate novel alternative uses for everyday objects (AUT—Alternate Uses Test; Guilford, 1967), and in another study they generated possible consequences of an unusual change in the world (e.g., living on without death) (Consequences Task; Torrance, 1962). The ESI, compared to a control induction, increased the number of appropriate alternate uses generated in the AUT and increased the number of appropriate consequences provided in the consequences task, thus providing clear evidence that episodic retrieval can contribute to divergent creative thinking.

Using the ESI to Identify Core Network Contributions to Constructive Episodic Processes

The constructive episodic simulation hypothesis suggests that a core network of brain regions supports imagining the future through episodic retrieval. In light of the behavioral evidence that ESI impacts episodic retrieval, if this manipulation increases recruitment of core network regions during future imagination relative to a control task, we have strong evidence that episodic retrieval based activity in the core network contributes to future simulation.

In an experiment by Madore, Szpunar, Addis, and Schacter (2016), participants viewed a series of object names. For each of these objects, participants either imagined a future event involving that object or generated a sentence about the object's size. Participants engaged in these tasks after receiving the ESI or a control induction in the scanner. The hippocampus, inferior parietal lobule,[1] and precuneus showed greater activation after the ESI than after the control induction for the future imagination task relative to the semantic control task, thereby indicating that the ESI impacts core network regions related to memory retrieval, as hypothesized. Several other regions outside of the core network showed similar effects.

Another combined fMRI-ESI study further supports the conclusion that episodic retrieval contributes to divergent thinking. In this experiment, participants performed a divergent thinking task (i.e., the AUT) and a control task that required generating object associates in the scanner. Hippocampal activity selectively increased during the AUT after the ESI compared to the control induction (Madore, Thakral, Beaty, Addis, & Schacter, 2019). Consistent with this finding, a related study revealed common engagement of the hippocampus when participants performed the AUT, remembered past experiences, and imagined future experiences (Beaty, Thakral, Madore, Benedek, & Schacter, 2018). Taken together with the previously reviewed behavioral evidence, these fMRI findings point towards a common neural underpinning for constructive uses of episodic retrieval that contribute to remembering, imagining, and divergent creative thinking.

Investigating Component Processes of Episodic Simulation Through Repetition Suppression

The common engagement of a core network for imagining the future and remembering the past suggests that regions within this network are involved in constructive episodic simulation. However, little is known about whether and how different regions within this core network contribute to these simulations. In this section, we explain how repetition suppression—the decrease in neural activity that is typically observed for a repeated stimulus in neuroimaging studies—can be used to isolate the brain regions that contribute to specific aspects of episodic simulation. We first highlight studies that have used repetition suppression to identify regions involved in simulating content-specific elements of episodic simulations (i.e., places, objects, people, and emotions). Then, we highlight studies that have used repetition suppression to identify regions involved in the flexible retrieval and restructuring of memories.

[1] Because many readers of this volume are likely interested in the temporoparietal junction and its role in theory of mind, and we discuss regions of activation within the inferior parietal cortex close to the TPJ, we provide additional anatomical details and loci of activation for such regions in our footnotes. Even with these additional details, some coordinates remain ambiguous. Due to anatomical variability between participants, the locations for these brain regions may vary across studies, so strong conclusions should be avoided in ambiguous cases (for further discussion on this topic, see the section *Neural substrates of theory of mind and episodic simulation*).

The Madore, Szpunar, et al. (2016) study reports a peak angular gyrus coordinate at (46, −68, 26). This location is posterior to the peak coordinate for the right TPJ, which is located at (56, −54, 20) according to a search for Theory of Mind in Neurosynth (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011).

Using Repetition Suppression to Identify Regions Involved in Simulating Content-Specific Elements of Episodic Simulation

By manipulating how often a specific component of a future simulation is repeated (e.g., asking participants to imagine a location three times or only once), researchers can test which brain regions show repetition-related differences in activity for the specific category that was repeated. If, for example, the parahippocampal place area shows repetition-related reductions in activity for a future simulation in which the cued location had been imagined three times (versus only once), then researchers can infer that the parahippocampal place area is involved in the simulation of locations.

Szpunar, St. Jacques, Robbins, Wig, and Schacter (2014) used this logic to identify component processes (location, person, and object simulation) of episodic simulations. They first asked participants to imagine a series of novel situations involving familiar people and locations. Participants were then presented with the same cues and asked to imagine the situation again. The third time that participants were asked to imagine scenarios, the cues were repeated or contained either a different person or location. By assessing differences in activity between the repeated and non-repeated trials, Szpunar, St. Jacques, et al. (2014) were able to identify regions sensitive to specific categories of stimuli. Among the regions sensitive to changes in location were the parahippocampal gyrus, retrosplenial cortex, precuneus, and ventral and medial prefrontal cortex. These regions significantly overlap with regions previously identified in location processing during other tasks (e.g., Epstein, 2008). For person repetitions, Szpunar, St. Jacques, et al. (2014) found differences in anterior and ventral medial prefrontal cortex, overlapping with regions previously found to be involved in person processing (e.g., Schurz, Radua, Aichhorn, Richlan, & Perner, 2014). These results, then, are consistent with the idea that the same cortical regions used for processing specific stimuli classes are re-used for imagining those classes of stimuli (Moulton & Kosslyn, 2009; Pearson & Kosslyn, 2015).

Szpunar, Jing, Benoit, and Schacter (2015) extended this repetition suppression strategy for identifying component processes of episodic simulations in order to distinguish between brain regions that support the simulation of positive, neutral, and negative future events. Although episodic future simulations in everyday life are frequently emotionally charged (D'Argembeau, Renaud, & Van der Linden, 2011), little is known about their neural underpinnings. In the study by Szpunar et al. (2015), participants were cued with novel combinations of familiar people, locations, and objects, and were instructed to imagine positive, negative, or neutral events involving those cues that could take place within the next 5 years. They typed a brief description of the imagined event to help them to re-imagine the same event the next day, when participants re-simulated half of the positive, negative, and neutral events. Szpunar et al. (2015) found that the pulvinar nucleus, a region previously linked with processing of aversive stimuli, showed repetition-related decreases in activity for frequently simulated negative events, whereas the orbitofrontal cortex, a region previously linked with processing of rewarding stimuli, showed repetition-related decreases for positive events. As in the experiment by Szpunar, St. Jacques, et al. (2014), this study provides additional evidence that brain regions involved in the processing of certain stimulus classes are recruited for imagining those stimulus classes as well.

Together, these studies show that repetition suppression paradigms can help to identify the regions involved in specific components of episodic simulation.

Using Repetition Suppression to Identify Regions Involved in the Flexible Retrieval and Restructuring of Memories

In addition to identifying regions involved in simulating specific types of content during future imagination, repetition suppression can also be used to study a central process of constructive simulations: the flexible retrieval and restructuring of information from memories. In the first of two studies addressing this general issue, St. Jacques, Szpunar, and Schacter (2017) adapted the repetition suppression paradigm from Szpunar, St. Jacques, et al. (2014) and used visual perspective shifting to study memory restructuring. Participants were cued with autobiographical memory prompts and were asked to retrieve memories from the original first-person (also termed "field perspective") or from a third-person observer's perspective. Participants retrieved each memory three times, either from the original or the new perspective. St. Jacques et al. (2017) found that the precuneus and the angular gyrus[2]—both regions within the core network that supports remembering and imagination—decreased their engagement during visual perspective shifting. These results were therefore interpreted as evidence that the precuneus and angular gyrus are involved in processes associated with perspective shifting and memory restructuring.

A second study investigated whether the constructive processes used for restructuring old events are shared with those used for imagining hypothetical events, and, if so, which brain regions support these processes. For this study, participants were again cued with autobiographical memory prompts and were asked to recall past experiences from the original viewpoint or from the observer perspective (St. Jacques, Carpenter, Szpunar, & Schacter, 2018). These trials were further divided into "veridical retrieval" and "episodic counterfactual thinking" conditions. During the counterfactual thinking task, participants were asked to imagine alternative ways an event could have occurred. The conjunction of counterfactual thinking > repeated memory and novel viewpoint > repeated memory revealed a common set of frontal regions involved in restructuring information during simulations.

[2] This study reports a peak angular gyrus coordinate at (46, −68, 26). This location is posterior to the peak coordinate for the right TPJ, which is located at (56, −54, 20) according to a search for Theory of Mind in Neurosynth (Yarkoni et al., 2011).

These regions—including ventrolateral prefrontal cortex, posterior dorsomedial prefrontal cortex, posterior dorsolateral prefrontal cortex, and posterior inferior parietal cortex[3]—overlap with the regions previously found to be more involved in episodic simulation than episodic retrieval (Benoit & Schacter, 2015). This finding is consistent with involvement of these regions in constructive processes, and their common engagement by two forms of memory restructuring suggests that repetition suppression can be used to identify brain regions involved in constructive processes.

In summary, then, repetition suppression has proven to be a useful tool for identifying which brain regions are involved in specific components of episodic simulation and memory, both when these components consist of specific content (such as locations, objects, emotions, and people), or constructive processes such as memory restructuring.

Manipulating Specific Regions Within the Core Network Supporting Memory and Imagination with TMS

The common network involved in remembering the past events and imagining the future has been identified through fMRI. Because fMRI is a correlational method, other tools such as transcranial magnetic stimulation (TMS) can be used to further test whether specific regions within this network are critical for remembering and imagining events. In one such study, Thakral, Madore, and Schacter (2017) used TMS to temporarily inhibit the left angular gyrus,[4] a region located within posterior parietal cortex and previously implicated in both episodic memory and future simulation through fMRI (Benoit & Schacter, 2015). Work with neuropsychological patients likewise suggests that the posterior parietal cortex might be causally involved in imagining detailed everyday scenarios: Berryhill, Picasso, Arnold, Drowos, and Olson (2010) found that patients with damage to this region generate less spatially coherent events with fewer entities, thoughts/emotions/actions, sensory descriptions, and spatial references when compared to control individuals. Conclusions from these patients, however, should be treated with some caution because neuropsychological patients often have diffuse damage that can influence behavior. As a result, additional evidence is needed to establish a critical role for the posterior parietal cortex in event constructions.

[3] This study reports two peak coordinates for the posterior inferior parietal cortex located at (−40, −64, 34) and (−42, −58, 26). It is unclear which label to apply to this region of activation. The peak coordinate for the left TPJ is located at (−54, −56, 22) according to a search for "Theory of Mind" in NeuroSynth (Yarkoni et al., 2011). The peak coordinate for the left angular gyrus is located at (−46, −66, 28) according to a search for "angular gyrus" in NeuroSynth (Yarkoni et al., 2011).

[4] The TMS target in this study was the left angular gyrus, located at (−48, −64, 30). This location is posterior to the peak coordinate for the left TPJ, located at (−54, −56, 22) according to a search for Theory of Mind in Neurosynth (Yarkoni et al., 2011).

Thakral et al. (2017) provided this complementary evidence. By inhibiting left angular gyrus activity during episodic simulation, episodic retrieval, and semantic control tasks, Thakral et al. tested whether this region is causally involved in episodic simulation and retrieval. After TMS was applied to the left angular gyrus (compared to a control vertex location), participants generated fewer episodic details (e.g., scene, people, action, and object details) when cued to remember personal past and future event, consistent with the hypothesized critical role of this region in simulation. By contrast, the imagined and remembered events contained a greater number of semantic details and commentary after left angular gyrus inhibition, suggesting that participants compensate for a decrease in episodic detail by relying on other types of information. Last, there was no detectable effect of TMS on the semantic control task, which involves the generation of object associates (e.g., *dog* for the cue *cat*).

These results suggest that the angular gyrus is indeed critically involved in both episodic retrieval and simulation and confirms that TMS can be used to identify regions on the cortical surface that play an important role in these processes.

Episodic Simulation, Memory, and Social Cognition

Episodic simulation, memory, and social cognition have been closely linked in the literature. Previous research suggests that the same neural circuits are used for both simulating future events and imagining the thoughts of others, but more recent research suggests that the two processes do not rely on a shared simulation architecture. We first discuss this work, then discuss research that instead proposes that episodic retrieval contributes to social cognition indirectly by providing individuals with access to memories and imagined situations that inform mental state judgments and can increase empathic responses.

Neural Substrates of Theory of Mind and Episodic Simulation

Early work on theory of mind, remembering the past, and imagining the future suggested that all three processes rely on the same brain network. A meta-analysis revealed neural overlap between the three types of thinking (Spreng, Mar, & Kim, 2009) and subsequent experiments found that BOLD trajectories in default network regions are similar for all three processes (Spreng & Grady, 2010; for replication, see DuPre, Luh, & Spreng, 2016).

Other work, however, suggests that the default network consists of three components rather than one. Andrews-Hanna, Reidler, Sepulcre, Poulin, and Buckner (2010) asked participants to think about scenarios involving themselves in the future, to think about themselves in the present, or to make semantic judgments about the future or present. A dorsal-medial network is preferentially engaged for self-referential processing (present self), while a medial temporal network becomes predominantly engaged for imagining a situation (future self). Two additional midline regions, consisting of medial prefrontal cortex and posterior cingulate cortex, appear to be involved in both processes (future self and present self). Consistent with the findings that two separate subsystems are involved in self-reference and episodic simulation, scene construction scores correlated highly with activity in the medial temporal network while affective self-referential composite scores correlated highly with activity in the dorsomedial network. Together, these results suggest that processes involving the thoughts of individuals are separable from those of episodic projection, which occur in the medial temporal lobe subsystem.

More recent studies suggest that these characterizations of the default network may be incomplete. Braga and Buckner (2017) and Braga, Van Dijk, Polimeni, Eldaief, and Buckner (in press) repeatedly scanned individuals to obtain more detailed maps of the default network than was possible in previous group-average studies. Functional connectivity revealed two interlocking networks, A and B, within what is classically defined as the default network. These two networks divide many of the regions previously implicated in both episodic memory and theory of mind, such as the posterior cingulate cortex, medial prefrontal cortex, and the inferior parietal lobule/temporoparietal junction. Regions within each network were more closely related than regions between these two networks, with some between-network functional connectivity correlations near zero. These findings suggest that some functional overlap between processes in the default network could simply be the result of blurring spatially adjacent networks. Because participants' brains have different shapes, neighboring networks will be in different locations for different individuals. So, when averaging together individual participants' maps to obtain a group-average map, blurring these smaller neighboring networks together makes them appear as one larger network (Braga & Buckner, 2017). Current research in the same laboratory is examining whether the functional overlap in episodic processing and theory of mind can be explained in part by these group-average map distortions (Lauren DiNicola, personal communication).

Contributions of Episodic Simulation and Memory to Social Cognition

While simulation of others' minds and simulation of the future may rely on separable mechanisms, episodic simulation and memory nonetheless contribute to social cognition. For example, people frequently draw on memories when making judgments about themselves and other people (Krienen, Tu, & Buckner, 2010) and replay social interactions to learn from them (Mar & Spreng, 2018). Participants further rely on episodic retrieval for social problem solving (Madore & Schacter, 2014). In addition, a growing line of work shows that episodic simulation may be especially beneficial for empathy.

Gaesser (2018) suggests that empathy and perspective taking benefit from episodic simulation when those we think about are not directly observable. For example, when we read a newspaper article and vividly imagine the situation of someone whose home was just flooded, we better appreciate the strain they are under and are more likely to feel empathy towards them. To test these hypotheses, Gaesser and Schacter (2014) asked participants to imagine helping another person, after which the participants were asked to rate how likely they were to help that person. In a separate condition, participants were asked to remember an episode in which they helped another person before rating the likelihood of helping them. When compared to several control conditions, including generating comments about how the person in the situation could be helped, participants reported greater willingness to help after imagining helping or remembering an event in which they had helped others. Further supporting the link between episodic simulation and empathy, Gaesser and Schacter (2014) observed that the sensory vividness of these imagined or remembered scenarios correlated with the degree of helping intentions.

In a subsequent study, Gaesser, Keeler, and Young (2018) directly manipulated scene imagery vividness by asking participants to imagine the helping situation in either a familiar context or an unfamiliar context and observed helping behavior in addition to the previously studied helping intentions. As expected, familiar contexts led to more vivid imagery, which led to greater helping intentions. In addition, familiar contexts led participants who were given money to allocate to themselves and another person to donate more of their money. Scene vividness in both familiar and unfamiliar context conditions was positively correlated with helping intentions and donation behavior. Again, these results suggest a role for episodic simulation in empathic behavior.

Importantly, this same experiment also examined the relationship between episodic simulation and mentalizing. After each trial, participants also rated the degree to which they took the perspective of the other person. In the familiar context condition, participants report greater perspective taking than in the unfamiliar context condition. Though limited by subjective ratings, these results are important because they indicate that scene imagery directly contributes to mentalizing. Gaesser (2018) suggests that this outcome further indicates that imagined situational information constrains the possible thoughts that a person may have.

Together, this body of work suggests that episodic simulation can be helpful for a variety of social cognitive tasks. While episodic simulation is not necessary for many of these tasks, as evidenced by amnesic individuals who are able to complete traditional theory of mind tasks (Rosenbaum, Stuss, Levine, & Tulving, 2007), episodic retrieval can nonetheless play an important role in social cognition because it allows people to richly imagine other individuals in specific contexts.

Concluding Comments

Episodic retrieval is important for far more than just remembering our past. Elements of our memories can be retrieved and flexibly recombined to imagine new events (e.g. Schacter & Addis, 2007b), and the resulting simulations contribute to a wide range of tasks, from personal planning to divergent thinking (e.g., Madore & Schacter, 2014; Madore et al., 2015; Spreng et al., 2010, 2015; for reviews, see Schacter, 2012, 2019; Schacter, Addis, & Szpunar, 2017). The constructive nature of episodic memory thus has significant implications for how we think about its functions.

In this chapter, we have illustrated three tools that researchers have leveraged to obtain a more detailed understanding of episodic simulation. First, the ESI has enabled researchers to test which tasks benefit from episodic retrieval and which regions of the brain mediate the contributions of episodic retrieval to these tasks (e.g., Madore, Szpunar, et al., 2016; Schacter & Madore, 2016). Second, repetition suppression has been successfully used to identify regions that are sensitive to specific component processes of simulation, such as the flexible recombination of memory elements (St. Jacques et al., 2018), or the imagination of new locations and objects (Szpunar, St. Jacques, et al., 2014). Third, TMS has been used to interfere with the normal functioning of a brain region within the core network to test which the region is critical for episodic simulation and retrieval (Thakral et al., 2017).

While more than a decade has passed since the early neuroimaging work on episodic simulation, many promising avenues of research remain open. First, we do not yet have a full understanding of the component processes that contribute to episodic simulation. The repetition suppression paradigms discussed earlier offer one promising path for future work to identify which brain regions are involved in additional components of simulations. Second, questions remain about the role that episodic retrieval plays in different types of future-oriented thinking. Szpunar, Spreng, and Schacter (2014) suggested that types of future thinking can be classified into four categories: simulation, prediction, intention setting, and planning. Within each of these categories, future-oriented thinking can range from fully semantic to richly episodic. An outstanding challenge for researchers will be to identify the situations in which different types of future thinking rely on episodic memory, semantic memory, or a combination of the two. Third, questions remain about the contributions of episodic simulation to tasks that are not traditionally associated with memory retrieval. For example, emotion regulation for worrisome events benefits from episodic specificity, as revealed by work using the ESI (e.g., Jing et al., 2016). Fourth, questions remain about the necessity of various regions within the core network for simulating events. By manipulating these regions with TMS, future research will likely further characterize which regions are critically involved in episodic simulation.

Finally, research that links episodic simulation with mentalizing and social cognition is just beginning, and much more needs to be done to characterize the extent and nature of those links (Gaesser, 2018; for related discussions on the link between episodic memory and theory of mind, see Laurita & Spreng, 2017). For example, it is unknown under which conditions episodic simulation contributes most to mentalizing. By adapting the paradigm used by Gaesser et al. (2018) and varying the types of situations and types of individuals participants are asked to engage with and measuring self-reported perspective taking, future research could start to address this issue.

To summarize, then, research motivated by the constructive episodic retrieval hypothesis has implicated episodic retrieval in many aspects of our everyday cognition. While challenges remain for fully understanding these processes, the three tools outlined in this chapter are aiding researchers in better understanding episodic simulation and its role in other forms of cognition.

Acknowledgements Preparation of this chapter was supported by grants from the National Institute of Mental Health (R01 MH060941) and National Institute on Aging (R01 AG008441) to DLS.

References

Addis, D. R., Musicaro, R., Pan, L., & Schacter, D. L. (2010). Episodic simulation of past and future events in older adults: Evidence from an experimental recombination task. *Psychology and Aging, 25*(2), 369–376.

Addis, D. R., Pan, L., Musicaro, R., & Schacter, D. L. (2016). Divergent thinking and constructing episodic simulations. *Memory, 24*(1), 89–97.

Addis, D. R., Pan, L., Vu, M. A., Laiser, N., & Schacter, D. L. (2009). Constructive episodic simulation of the future and the past: Distinct subsystems of a core brain network mediate imagining and remembering. *Neuropsychologia, 47*(11), 2222–2238.

Addis, D.R., Wong, A.T. & Schacter, D.L. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. Neuropsychologia, 45, 1363–1377.

Addis, D. R., Wong, A. T., & Schacter, D. L. (2008). Age-related changes in the episodic simulation of future events. *Psychological Science, 19*(1), 33–41.

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., & Buckner, R. L. (2010). Functional-anatomic fractionation of the brain's default network. *Neuron, 65*(4), 550–562.

Beaty, R. E., Benedek, M., Silvia, P. J., & Schacter, D. L. (2016). Creative cognition and brain network dynamics. *Trends in Cognitive Sciences, 20*(2), 87–95.

Beaty, R. E., Thakral, P. P., Madore, K. P., Benedek, M., & Schacter, D. L. (2018). Core network contributions to remembering the past, imagining the future, and thinking creatively. *Journal of Cognitive Neuroscience, 30*(12), 1939–1951.

Benoit, R. G., & Schacter, D. L. (2015). Specifying the core network supporting episodic simulation and episodic memory by activation likelihood estimation. *Neuropsychologia, 75*, 450–457.

Berryhill, M. E., Picasso, L., Arnold, R., Drowos, D., & Olson, I. R. (2010). Similarities and differences between parietal and frontal patients in autobiographical and constructed experience tasks. *Neuropsychologia, 48*(5), 1385–1393.

Braga, R. M., & Buckner, R. L. (2017). Parallel interdigitated distributed networks within the individual estimated by intrinsic functional connectivity. *Neuron, 95*(2), 457–471.

Braga, R. M., Van Dijk, K. R. A., Polimeni, J. R., Eldaief, M. C., & Buckner, R. L. (in press). Parallel distributed networks resolved at high resolution reveal close juxtaposition of distinct regions. *Journal of Neurophysiology, 121*, 1513. https://doi.org/10.1152/jn.00808.2018

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network. *Annals of the New York Academy of Sciences, 1124*(1), 1–38.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences, 11*(2), 49–57.

Carpenter, A. C., & Schacter, D. L. (2017). Flexible retrieval: When true inferences produce false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(3), 335–349.

Carpenter, A. C., & Schacter, D. L. (2018). Flexible retrieval mechanisms supporting successful inference produce false memories in younger but not older adults. *Psychology and Aging, 33*(1), 134–143.

D'Argembeau, A., Renaud, O., & Van der Linden, M. (2011). Frequency, characteristics and functions of future-oriented thoughts in daily life. *Applied Cognitive Psychology, 25*(1), 96–103.

De Brigard, F., Addis, D. R., Ford, J. H., Schacter, D. L., & Giovanello, K. S. (2013). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia, 51*(12), 2401–2414.

Dede, A. J., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2016). Autobiographical memory, future imagining, and the medial temporal lobe. *Proceedings of the National Academy of Sciences, 113*(47), 13474–13479.

Duff, M. C., Kurczek, J., Rubin, R., Cohen, N. J., & Tranel, D. (2013). Hippocampal amnesia disrupts creative thinking. *Hippocampus, 23*(12), 1143–1149.

DuPre, E., Luh, W. M., & Spreng, R. N. (2016). Multi-echo fMRI replication sample of autobiographical memory, prospection and theory of mind reasoning tasks. *Scientific Data, 3*, 160116.

Epstein, R. A. (2008). Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends in Cognitive Sciences, 12*(10), 388–396.

Fisher, R. P., & Geiselman, R. E. (1992). *Memory enhancing techniques for investigative interviewing: The cognitive interview*. Springfield, IL: Charles C. Thomas Publisher.

Gaesser, B. (2018, November 7). Episodic mind reading: Mentalizing guided by imagined and remembered scenes. *PsyArXiv*.

Gaesser, B., Keeler, K., & Young, L. (2018). Moral imagination: Facilitating prosocial decision-making through scene imagery and theory of mind. *Cognition, 171*, 180–193.

Gaesser, B., Sacchetti, D. C., Addis, D. R., & Schacter, D. L. (2011). Characterizing age-related changes in remembering the past and imagining the future. *Psychology and Aging, 26*(1), 80–84.

Gaesser, B., & Schacter, D. L. (2014). Episodic simulation and episodic memory can increase intentions to help others. *Proceedings of the National Academy of Sciences, 111*(12), 4415–4420.

Gerlach, K. D., Spreng, R. N., Gilmore, A. W., & Schacter, D. L. (2011). Solving future problems: Default network and executive activity associated with goal-directed mental simulations. *NeuroImage, 55*(4), 1816–1824.

Guilford, J. P. (1967). *The nature of human intelligence*. New York, NY: McGraw Hill.

Gilbert, D. T., & Wilson, T. D. (2007). Prospection: Experiencing the future. *Science, 317*(5843), 1351–1354.

Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences, 104*(5), 1726–1731.

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in Cognitive Sciences, 11*(7), 299–306.

Ingvar, D. H. (1985). "Memory of the future": An essay on the temporal organization of conscious awareness. *Human Neurobiology, 4*(3), 127–136.

Jing, H. G., Madore, K. P., & Schacter, D. (2016). Worrying about the future: An episodic specificity induction impacts problem solving, reappraisal, and well-being. *Journal of Experimental Psychology: General, 145*(4), 402–418.

Jing, H. G., Madore, K. P., & Schacter, D. L. (2017). Preparing for what might happen: An episodic specificity induction impacts the generation of alternative future events. *Cognition, 169*, 118–128.

Krienen, F. M., Tu, P. C., & Buckner, R. L. (2010). Clan mentality: Evidence that the medial prefrontal cortex responds to close others. *Journal of Neuroscience, 30*(41), 13906–13915.

Laurita, A. C., & Spreng, R. N. (2017). The hippocampus and social cognition. In *The hippocampus from cells to systems* (pp. 537–558). Cham, Switzerland: Springer.

Levine, B., Svoboda, E., Hay, J. F., Winocur, G., & Moscovitch, M. (2002). Aging and autobiographical memory: Dissociating episodic from semantic retrieval. *Psychology and Aging, 17*(4), 677.

Madore, K. P., Addis, D. R., & Schacter, D. L. (2015). Creativity and memory: Effects of an episodic-specificity induction on divergent thinking. *Psychological Science, 26*(9), 1461–1468.

Madore, K. P., Gaesser, B., & Schacter, D. L. (2014). Constructive episodic simulation: Dissociable effects of a specificity induction on remembering, imagining, and describing in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(3), 609–622.

Madore, K. P., & Schacter, D. L. (2014). An episodic specificity induction enhances means-end problem solving in young and older adults. *Psychology and Aging, 29*(4), 913–924.

Madore, K. P., & Schacter, D. L. (2016). Remembering the past and imagining the future: Selective effects of an episodic specificity induction on detail generation. *The Quarterly Journal of Experimental Psychology, 69*(2), 285-298.

Madore, K. P., Szpunar, K. K., Addis, D. R., & Schacter, D. L. (2016). Episodic specificity induction impacts activity in a core brain network during construction of imagined future experiences. *Proceedings of the National Academy of Sciences, 113*(38), 10696–10701.

Madore, K. P., Thakral, P. P., Beaty, R. E., Addis, D. R., & Schacter, D. L. (2019). Neural mechanisms of episodic retrieval support divergent creative thinking. *Cerebral Cortex, 29*(1), 150–166.

Mar, R. A., & Spreng, R. N. (2018). Episodic memory solves both social and nonsocial problems, and evolved to fulfill many different functions. *Behavioral and Brain Sciences, 41*, e20.

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., & Macrae, C. N. (2007). Wandering minds: The default network and stimulus-independent thought. *Science, 315*(5810), 393–395.

Moulton, S. T., & Kosslyn, S. M. (2009). Imagining predictions: Mental imagery as mental emulation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1521), 1273–1280.

Ochsner, K. N., Knierim, K., Ludlow, D. H., Hanelin, J., Ramachandran, T., Glover, G., & Mackey, S. C. (2004). Reflecting upon feelings: An fMRI study of neural systems supporting the attribution of emotion to self and other. *Journal of Cognitive Neuroscience, 16*(10), 1746–1772.

Okuda, J., Fujii, T., Ohtake, H., Tsukiura, T., Tanji, K., Suzuki, K., … Yamadori, A. (2003). Thinking of the future and past: The roles of the frontal pole and the medial temporal lobes. *NeuroImage, 19*(4), 1369–1380.

Pearson, J., & Kosslyn, S. M. (2015). The heterogeneity of mental representation: Ending the imagery debate. *Proceedings of the National Academy of Sciences, 112*(33), 10089–10092.

Platt, J. J., & Spivack, G. (1975). *Manual for the mean-end problem-solving procedure (MEPS): A measure of interpersonal cognitive problem-solving skill*. Philadelphia, PA: Hahnemann Community Mental Health/Mental Retardation Center, Hahnemann Medical College and Hospital.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10*(2), 59–63.

Race, E., Keane, M. M., & Verfaellie, M. (2011). Medial temporal lobe damage causes deficits in episodic memory and episodic future thinking not attributable to deficits in narrative construction. *Journal of Neuroscience, 31*(28), 10262–10269.

Raichle, M. E. (2015). The brain's default mode network. *Annual Review of Neuroscience, 38*, 433–447.

Rosenbaum, R. S., Stuss, D. T., Levine, B., & Tulving, E. (2007). Theory of mind is independent of episodic memory. *Science, 318*(5854), 1257.

Schacter, D. L. (2012). Adaptive constructive processes and the future of memory. *American Psychologist, 67*(8), 603–613.

Schacter, D. L. (2019). Implicit memory, constructive memory, and imagining the future: A career perspective. *Perspectives on Psychological Science, 14*, 256–272.

Schacter, D. L., & Addis, D. R. (2007a). Constructive memory: The ghosts of past and future. *Nature, 445*, 27.

Schacter, D. L., & Addis, D. R. (2007b). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362*(1481), 773–786.

Schacter, D. L., & Addis, D. R. (2020). Memory and imagination: Perspectives on constructive episodic simulation. In A. Abraham (Ed.), *The Cambridge Handbook of the Imagination* (pp. 111-131). Cambridge, MA: Cambridge University Press.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience, 8*(9), 657–661.

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The future of memory: Remembering, imagining, and the brain. *Neuron, 76*(4), 677–694.

Schacter, D. L., Benoit, R. G., De Brigard, F., & Szpunar, K. K. (2015). Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions. *Neurobiology of Learning and Memory, 117*, 14–21.

Schacter, D. L., Guerin, S. A., & St. Jacques, P. L. (2011). Memory distortion: An adaptive perspective. *Trends in Cognitive Sciences, 15*(10), 467–474.

Schacter, D. L., & Madore, K. P. (2016). Remembering the past and imagining the future: Identifying and enhancing the contribution of episodic memory. *Memory Studies, 9*(3), 245–255.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews, 42*, 9–34.

Sheldon, S., McAndrews, M. P., & Moscovitch, M. (2011). Episodic memory processes mediated by the medial temporal lobes contribute to open-ended problem solving. *Neuropsychologia, 49*(9), 2439–2447.

Spreng, R. N., Gerlach, K. D., Turner, G. R., & Schacter, D. L. (2015). Autobiographical planning and the brain: Activation and its modulation by qualitative features. *Journal of Cognitive Neuroscience, 27*(11), 2147–2157.

Spreng, R. N., & Grady, C. L. (2010). Patterns of brain activity supporting autobiographical memory, prospection, and theory of mind, and their relationship to the default mode network. *Journal of Cognitive Neuroscience, 22*(6), 1112–1123.

Spreng, R. N., Mar, R. A., & Kim, A. S. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience, 21*(3), 489–510.

Spreng, R. N., Stevens, W. D., Chamberlain, J. P., Gilmore, A. W., & Schacter, D. L. (2010). Default network activity, coupled with the frontoparietal control network, supports goal-directed cognition. *NeuroImage, 53*(1), 303–317.

Squire, L. R., van der Horst, A. S., McDuff, S. G., Frascino, J. C., Hopkins, R. O., & Mauldin, K. N. (2010). Role of the hippocampus in remembering the past and imagining the future. *Proceedings of the National Academy of Sciences, 107*(44), 19044–19048.

St. Jacques, P. L., Carpenter, A. C., Szpunar, K. K., & Schacter, D. L. (2018). Remembering and imagining alternative versions of the personal past. *Neuropsychologia, 110*, 170–179.

St. Jacques, P. L., Szpunar, K. K., & Schacter, D. L. (2017). Shifting visual perspective during retrieval shapes autobiographical memories. *NeuroImage, 148*, 103–114.

Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences, 30*(3), 299–313.

Szpunar, K. K., Jing, H. G., Benoit, R. G., & Schacter, D. L. (2015). Repetition-related reductions in neural activity during emotional simulations of future events. *PLoS One, 10*(9), e0138354.

Szpunar, K. K., Spreng, R. N., & Schacter, D. L. (2014). A taxonomy of prospection: Introducing an organizational framework for future-oriented cognition. *Proceedings of the National Academy of Sciences, 111*(52), 18414–18421.

Szpunar, K. K., St. Jacques, P. L., Robbins, C. A., Wig, G. S., & Schacter, D. L. (2014). Repetition-related reductions in neural activity reveal component processes of mental simulation. *Social Cognitive and Affective Neuroscience, 9*(5), 712–722.

Szpunar, K. K., Watson, J. M., & McDermott, K. B. (2007). Neural substrates of envisioning the future. *Proceedings of the National Academy of Sciences, 104*(2), 642–647.

Schacter, D. L., Addis, D. R., & Szpunar, K. K. (2017). Escaping the past: Contributions of the hippocampus to future thinking and imagination. In The hippocampus from cells to systems (pp. 439–465). Springer, Cham.

Thakral, P. P., Madore, K. P., & Schacter, D. L. (2017). A role for the left angular gyrus in episodic simulation and memory. *Journal of Neuroscience, 37*(34), 8142–8149.

Torrance, P. E. (1962). *Guiding creative talent*. Englewood Cliffs, NJ: Prentice-Hall.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26*(1), 1–12.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53*, 1–25.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods, 8*(8), 665–670.

Proactive by Default

Shira Baror, Elissa M. Aminoff, and Moshe Bar

The Brain's Default Mode Is Not Rest

By default, indeed, the mind wanders. The world of "mind wandering" research rapidly expanded with the discovery of the "default mode network" (DMN), a set of brain regions found to be consistently more activated when one is free of an experimental task (Raichle et al., 2001; Shulman et al., 1997). This network primarily includes the posterior cingulate cortex (PCC), the medial prefrontal cortex (mPFC), the lateral parietal cortex, and the hippocampal formation (including the entorhinal cortex and parahippocampal cortex) (Bar, Aminoff, Mason, & Fenske, 2007; Buckner, Andrews-Hanna, & Schacter, 2008). Rather than displaying sparse or noisy activations, the human neural "baseline" proved to be well-organized, consistently active, and to perform what is suggested to be fundamental processing.

With the remarkable evidence of the brain's robust and synchronized circuits at "rest" came a no less remarkable number of attempts to fathom this network's primary function. Various accounts have been raised, each supported by a body of research. Mentalizing (Andrews-Hanna, 2012), self-referential processing (Buckner & Carroll, 2007), stimulus-independent thought (Schooler et al., 2011), memory of the past and envisioning of the future (Addis, Pan, Vu, Laiser, & Schacter, 2009), creative thinking (Baird et al., 2012), spontaneous cognition (Andrews-Hanna, Reidler, Huang, & Buckner, 2010), prospection and theory of mind (Spreng, Mar, & Kim, 2009) and the generation of predictions (Bar, 2007) have all been associated with DMN activation. This exceptionally large number of theories has naturally yielded a corresponding number of debates: is the default activity

S. Baror · M. Bar (✉)
The Gonda Multidisciplinary Brain Research Center, Bar-Ilan University, Ramat Gan, Israel
e-mail: moshe.bar@biu.ac.il

E. M. Aminoff
Department of Psychology, Fordham University, Bronx, NY, USA
e-mail: eaminoff@fordham.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_23

stimulus-dependent or stimulus-independent, is it effortless or resource-consuming, is it triggered only spontaneously or also by intentional means, is its activation in contrast with other task-related neural networks or do co-activations occur, to name a few.

To be involved in such myriad of processes, the function of this network ought to be more fundemental, one that provides the building block to all other processes with which this network has been implicated. We suggest this function to be the proactive activation of contextual associations (Bar, 2004; Bar, 2007; Bar & Aminoff, 2003; Bar et al., 2007; Fenske, Aminoff, Gronau, & Bar, 2006; Kveraga, Ghuman, & Bar, 2007).

In this chapter, we review our overarching proposal that the default activity of the brain represents the activation of context-based associations. Such associative activations occur continuously during "rest" and sit at the basis of mentalizing, planning, simulating as well as the other processes with which the DMN has been linked, providing the basis for their realization. We demonstrate that across domains and situations, the extraction and activation of associative information forms the primary criterion for evoking activity in this network.

To support our proposal, we discuss findings that show that this neural network is not only the intrinsic "resting state" neural signature of internal mentation but it is also activated when one is engaged with external information that involves contextual processing. We discuss the function of associations in mentalizing and examine the extent to which other functions that have been attributed to this network rely on contextually associative processes.

Proactive over Rest: The Contextual Associations Network

In parallel to the accumulating evidence regarding the DMN and its functions, we have noticed that our own fMRI studies on contextual associations elicit activation maps that overlap with the DMN (Bar & Aminoff, 2003). By asking participants to perform a visual recognition task of objects that elicit either strong contextual information (e.g., a traffic light, which specifically suggests the context of a street junction) or weak contextual information (e.g., a light bulb, which can be found in many possible contexts), we were able to identify a set of brain regions that is sensitive to the availability of contextual information. This network, mediating context-based associations, comprises the parahippocampal cortex (PHC), the retrosplenial complex (RSC), and the medial prefrontal cortex (mPFC) (Aminoff, Gronau, & Bar, 2007; Aminoff, Kveraga, & Bar, 2013; Aminoff & Tarr, 2015; Bar & Aminoff, 2003; Bar et al., 2007). The retrosplenial complex refers to the medial parietal cortex, which includes the retrosplenial cortex, the posterior cingulate, and the anterior portions of the precuneus. Typically, contextual associations elicited activity that was in more anterior portions of the medial parietal cortex, focused on the retrosplenial cortex, but also extended to posterior regions such as the posterior cingulate and the precuneus. The striking overlap of this contextual associations network with the DMN is evident and reproduced in Fig. 1.

The notable overlap between the cortical network mediating contextual associations and the default network is not a coincidence. We argue that all those diverse cognitive processes that are linked with the operation of the DMN, from self-referential thought to mental simulations, planning and theory of mind, rely on associations at their core, and thus all recruit similar foundational circuitry. The overlap between the contextual associations network and the DMN provides the platform for a parsimonious theory that would explain the functional overlap between all those diverse processes that seem to reside in the DMN.

Outside the lab, the artificial boundaries between social, emotional, perceptual, or cognitive processes become vague as all these domains intertwine in continuous streams of thought. For example, while waiting for a train en route to a job interview, thoughts that occur can range from and mix within domains such as goal orienting toward searching the train and boarding it, planning what to say when we arrive at the interview, anticipating the interviewer's questions, calming yourself down, and rehearsing your to-do list for when the interview ends. Clearly, our brain is not passive, simply reacting to information coming from the senses, but rather it is proactively anticipating future events (Bar, 2009a). To generate predictions proactively, the underlying psychological and neural mechanisms rely on the constant activation of relevant associations.

In the past decade, many studies demonstrated the involvement of context-based associations in various processes. In the visual domain, beyond being activated during recognition of objects containing strong contextual information, the contextual associations network was active when processing scenes (Bar, Aminoff, & Schacter, 2008) as well as when processing famous faces (Bar, Aminoff, & Ishai, 2008), which similarly elicit strong contexts. Additionally, activity in the contextual associations network, elicited by strong contextual information, generalizes over different viewpoints (Marchette, Vass, Ryan, & Epstein, 2015), supporting the notion of a contextualy integrated representation from direct sensory input (Cheung & Bar, 2014). The temporal course of action of contextual information takes place very early in the perceptual process, suggested to facilitate recognition by the generation of context-based predictions (Kveraga et al., 2011). Taken together, these findings support the critical role associative information plays in interpreting visual information. Past experiences include affective associations as well, such that social and emotional information acquired previously serve as predictive context for upcoming events (Tamir & Thornton, 2018). Extracting such cues that rely on contextual processing has been suggested to be critical for social cognition (see Barrett & Bar, 2009; Brown & Brune, 2012; Schilbach, Eickhoff, Rotarska-Jagiela, Fink, & Vogeley, 2008) and we will elaborate on this point further in the section dedicated to contextual associations and mentalizing.

In sum, the network of cortical regions activated by contextual associations largely overlap with the DMN. In upcoming sections, we will demonstrate how all those other functions with which the DMN has been implicated are based on such associative activation, and therefore all activate the same network.

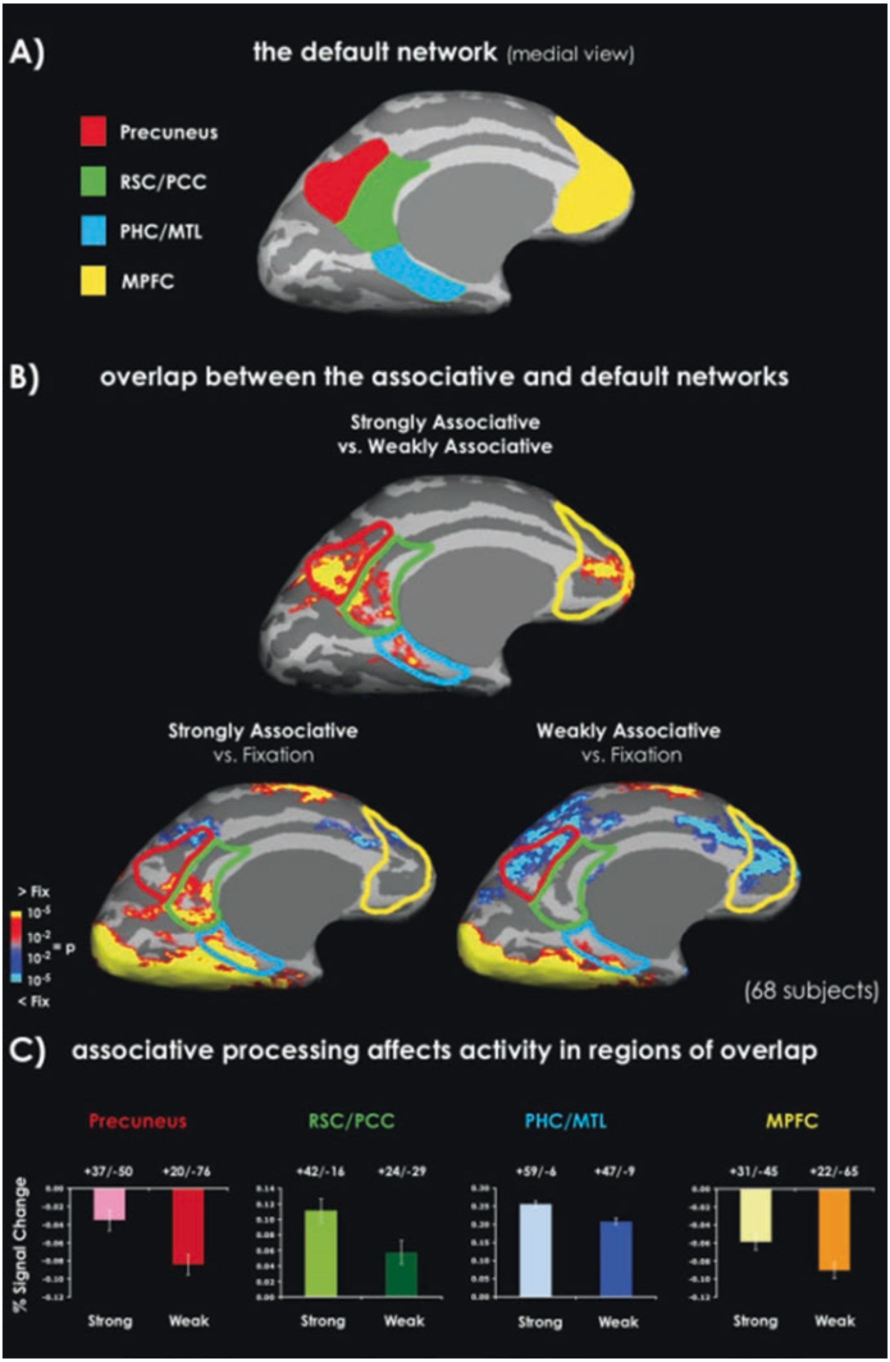


Fig. 1 The overlap between the contextual associations network and the default mode network. (**a**) Medial view illustration of the core DMN nodes. (**b**) The overlap in brain regions that are sensitive to processing strong vs. weak associations and the DMN. (**c**) The specific effect of associative processing on fMRI activity in each of the overlapping regions. (Figure is reproduced, with permission, from Bar et al., 2007)

Associative Processing Explains Default Mental Operations

The DMN was defined by the areas that were found to be *more active during rest* than during task, and thus were, somewhat poorly, named "deactivations." The definition of the contextual associations network, on the other hand, is the areas that are *more active during task* than during rest, and thus are taken as "activations." It is the same network, only in studies of contextual associations this network is recruited during task, and in studies of the DMN this network is recruited during rest.

Given the overlap between the two networks, the direction of attention (inward vs. outward) and the difference between rest and on-task settings cannot account for the activation of the DMN. Instead, a more careful look at most studies reveals that the activation of contextual associations can serve as a more plausible alternative, and a more generalized framework. In other words, *our logic has been that because our experiments with contextual associations yielded activation in the same areas that are otherwise active when people are not engaged in a task, activation of contextual associations best describes the mental operation that takes place during rest.* Everything else builds upon it.

It is important to note that default operations have also been attributed to episodic memory and to semantic conceptualization, as detailed in other chapters in this book. We propose that associations are the elementary unit of memory and thought, and therefore these alternative accounts commonly elicit associative processing in the DMN.

Associative "Off-Task"

Naturally, the default tendency of any system is found when exploring it unconstrained. This is probably why historically the default mode of the brain was dubbed "resting state" (Raichle et al., 2001; Shulman et al., 1997). When all demands are withheld, the brain spontaneously engages in internal processing of mind wandering (Mason et al., 2007), often discussed in terms of "perceptual decoupling" in the sense that internal processes during mind wandering are detached from immediate percepts and thus are stimulus-independent, or "task-unrelated thought" in the sense that they are spontaneous and not related to a given task (Schooler et al., 2011).

The studies that found DMN activity in the form of "deactivations" compared with task-related activity used tasks that do not engage associative information [e.g., go/no go task (Christoff, Gordon, Smallwood, Smith, & Schooler, 2009); working memory tasks (Mason et al., 2007); visual attention (Raichle et al., 2001)]. The "on-task"/"off-task" terminology with regard to the DMN has therefore originated from the difference in activation between passive fixation during the "resting" baseline, and tasks that primarily rely on the dorsal attention network, a set of regions that consequently was shown to be anti-correlated with the DMN (Fox et al., 2005). Therefore, the definition of the default mode may be a confound of the contrast used rather than the more fundamental cognitive mechanisms that this mode underlies. When people are at “rest” in the condition typically used to define the default mode, people tend to think about their past, their future, or experience the flow of thoughts. Such spontaneous processes, beyond being internally driven, necessitate the activation of already stored associations of various kinds, which fit well with our proposal that the elemental operation carried out by the DMN is the activation of contextual associations.

Associative “On-Task”

The fact that a default state of a system is unveiled by means of an unimpeded exploration does not exclude the possibility to prompt the same state by employing experimental manipulations in the quest to tap its exact underlying operation. As shown by numerous studies afterwards, the DMN actively supports many cognitive processes that take place while being on-task as well (e.g., Addis et al., 2009; Axelrod, Rees, Lavidor, & Bar, 2015; Bar et al., 2007; Hassabis, Kumaran, & Maguire, 2007; Sestieri, Shulman, & Corbetta, 2017). Because these previous studies have driven the activity in the DMN above baseline, it follows that the tasks involved that lead to such results may be more astutely targeting what these regions are processing. Such experiments demonstrate that the processes that are found to engage our minds during rest can be experimentally invoked and examined, confirming that merely being “off-task” cannot fully characterize default activity.

Most of those tasks that were shown to invoke the DMN circuitry rely on associative activation. Be it mind wandering about the past (Fox, Spreng, Ellamil, Andrews-Hanna, & Christof, 2015), engaging in autobiographical (Baird, Smallwood, & Schooler, 2011; Spreng et al., 2009), or episodic memory (Addis et al., 2009) all of these lean on previously acquired information as their platform. As we therefore suggested, while DMN activation is traditionally identified by contrasting attentional tasks with either rest or tasks that rely on internal processing, the availability and involvement of associations in these contrasts should be more explicitly regarded.

Recent evidence shows that even when stripped of their meaningful content, associations activate core nodes of the DMN. Specifically, Aminoff et al. (2007), and later, Aminoff and Tarr (2015) demonstrated how processing meaningless items that share their location or identity, and thus form spatial and identity associations, activate the primary regions of the contextual associations network, namely the PPA, the RSC, and also a non-DMN region, the occipital place area (OPA). Moreover, the activity in these regions correlated with how well participants learned these associations. For example, the better an association was learned between meaningless shapes, the more signal was elicited from the RSC when comparing associative conditions to non-associative conditions (Aminoff & Tarr, 2015). These findings substantiate our main claim that this network not only shows diminished deactivations during “rest” but is increasingly activated in the presence of associative information. Most notably, even when meaningless in content, the presence of associative information is sufficient to elicit activation in this neural network.

To further support this argument, one would predict that even within a given task involving internally guided thought, neural activity would change as a function of the level of engagement in associations. Evidence supporting this proposition is found in an fMRI study conducted by Gilmore, Nelson, and McDermott (2016), who were able to control the availability of contextual information by asking participants to both remember personal episodic past events and imagine possible ones. In their experiment, easier access to contextual associations was assumed for remembered compared with imagined events, and the researchers found greater activation of the contextual associations network (primarily the PHC and the RSC) in remembered versus imagined events. This finding implies that when attention is decoupled from sensory stimulations, the neural activation in the network still changes as a function of the availability of contextual information. Similarly, Tamir and Mitchell (2011) found greater activation of this network when thinking about proximal compared with distal events, suggesting that these brain regions are sensitive to the availability of rich and immediate contextual associations.

To conclude, DMN patterns of activation seem to defy the dichotomous distinction between off- and on-task settings. Instead, the evidence highlight that this network is primarily occupied with how associative the processes are, above and beyond the specific settings in which they occur.

The Mentalizing Train Rides Associative Tracks

Mentalizing has been strongly correlated with DMN processing. Therefore, after showing that the DMN is predominantly triggered by contextual associative processing, we turn to discuss the involvement of these contextual associative activations in mentalizing.

Mentalizing relies on the ability to represent information that is not conveyed by one's current sensory experience. It is a high-level component in human cognition that allows the continuous sense of consciousness, bridging the here and now with the past and the future (Andrews-Hanna, 2012). Mentalizing allows both inferring the behavior and mental states of others and simulating one's own personal choice-alternatives and actions. In fact, as will be elaborated, we claim that mentalizing primarily allows the understanding of others via associating the other person's state with our own. The "mentalizing network" corresponds with the DMN (Frith & Frith, 2006), with the focus on the PCC, as well as the mPFC, and the temporal poles, brain regions that overlap with the contextual associations network as well.

This overlap is not surprising: from the contextual association's perspective, mentalizing involves the activation of personal experiences and associations in at least two ways. First, mentalizing involves extracting agent-related associations. Similarities, in the form of associations, between the behavior of others and our own actions in analogous past experiences may be used as a form of information that allows us to personally relate to the information that is being conveyed and by that extend its possible implications (Spreng & Mar, 2012). For example, it is easier for us to relate to a friend's professional challenge when that friend is a colleague, compared with when that friend's occupation is completely different than ours. Being able to draw associations between their experience and ours allows us to put ourselves in their shoes. If this is true, the more associations one can draw between their own and another person's actions, the more the DMN activity will be observed during metalizing.

The second manner in which mentalizing involves associative information pertains to context-related associations, namely, to the context in which the information conveyed by others is understood. Associating one's previous experience to the current social setting is utilized as context for what would be socially appropriate in a given situation. For example, one's crying behavior is differently understood when that person had just married the love of their life (i.e., the context of a wedding), or alternatively if they are mourning the loss of a close friend (i.e., the context of a funeral). As such, associative information becomes the point of reference for inference (Frith & Frith, 2006), not only in understanding others, but in judging the appropriateness of their actions as well. If this is true, familiarity with the social context in which mentalizing takes place is expected to manifest in the ability to evaluate other's behavior, and more importantly, to be represented in increased DMN activation.

Both associative aspects in mentalizing highlight how DMN activity is expected to not only underlie the categorical function one is executing (i.e., mentalizing) but rather to be sensitive to gradual changes of associative content within that process. In line with the predictions made above, activation in one's mentalizing network has been found to be sensitive to the level of similarity to the subject of inference (Mitchell, Macrae, & Banaji, 2006). In these studies, participants were asked to mentalize the opinions of similar and dissimilar others while undergoing fMRI scan. The results show that the mPFC (a sub-region shared by the mentalizing network, DMN, and contextual associations network), and specifically its ventral region, is more strongly engaged when making judgements about the mental state of similar rather than dissimilar others, implying that extracting the similarity between the subject of inference (in this case, the mental state of another) and our personal relevant experience is underlined in the network's activation. In other words, DMN activity diverges between two mentalizing processes that are different in their associative availability. Another line of work shows that activity in the mPFC correlates with levels of stereotyping and prejudice, behaviors that strongly rely on associations between physical properties and assumed traits (Amodio, 2014).

In another study, differences in hippocampal activity between understanding similar and dissimilar others have been found selectively when autobiographical memories of real past events (rather than hypothetical ones) were engaged by the mentalizing participants (Perry, Hendler, & Shamay-Tsoory, 2011). This finding points to the interaction between the two ways we suggest that associations mediate mentalizing: extraction of agent-related associations (i.e., similarity of the mentalizing person to the inferred individual) and context-related associations (i.e., similarity of the social setting to real-life past experiences).

The two associative aspects in mentalizing (i.e., agent-related and context-related associations) correspond with the object-based and context-based mechanisms that we previously claimed to contribute to top-down facilitation of object recognition, within the framework of the contextual associations network (Fenske et al., 2006). While in the context of object recognition we characterized specific mechanisms for each of these associative processes, it would be interesting to further explore how in the social domain, agent-related and context-related associations jointly facilitate social behavior.

Further compelling evidence for a more direct involvement of memory-related processes in mentalizing comes from a study in single-neuron recording of human brain activity (Mukamel, Ekstrom, Kaplan, Iacoboni, & Fried, 2010). Recording single-neuron activity during self-execution of gestures and observation of others executing these actions revealed a unique pattern of activation in brain regions implicated with the contextual associations network. Unlike cells in most brain regions that were sensitive to one action aspect (either execution or observation), a selective group of cells found in the hippocampus and in the PHC responded to both action aspects. Under the framework of contextual associations, this finding adds support to the idea that computations carried out in these regions bind the inference of others with self-related experience. And although multiple mechanisms can be envisioned for what content exactly is represented by these regions' neuronal activity, their associative nature is well demonstrated in the results.

Taken together, these findings bolster the idea that mentalizing requires the activation of contextual associations, which can then be utilized for generating predictions regarding the traits and intentions of others around us (for further review, see Brown & Brune, 2012).

To summarize, inferring the mental state of others could have theoretically relied solely on the actual information these people explicitly convey while putting self-experience and prior assumptions aside. However, the findings cited above imply otherwise. As with making sense of our sensory experiences, which involves top-down processes of utilizing prior knowledge (Bar et al., 2006), it is suggested that mentalizing involves the simulation of social contextual information as well (Tamir, Bricker, Dodell-Feder, & Mitchell, 2015). It is a private case of inference processes, and as such, be it visual, auditory, emotional or social information, one associates this information by asking "what is this *like*," rather than simply "what is it" to understand and be able to relate to it (Bar, 2009a). So, whether we try to find our date amongst the crowd (i.e., a visual recognition process) or figure out their feelings once we have met them (i.e., a social mentalization process), we utilize contextual associations, and generate predictions accordingly, using the DMN.

Task Demands and Spontaneity in the Activation of Contextual Associations

In the previous sections, we showed that associative activation forms the cornerstone of mentalizing, as well as many other functions implicated with the DMN activity. We now turn to address the main debates concerning the nature of processing that have been linked to default activity. We further solidify associative activation as the framework under which we suggest these questions should be resolved.

How resource-consuming is the default and proactive tendency to internally activate associations? Does this process's spontaneous mode contradict a more intentional form of activation? These yet unresolved debates regarding cognitive spontaneity and cognitive cost in default activation give rise to critical predictions about the role this human tendency play in our daily life.

On the one hand, it has been previously claimed that the DMN deactivates as task demands increase (Greicius, Krasnow, Reiss, & Menon, 2003). This claim was supported by the anticorrelation found between the DMN showing task-related deactivations (i.e., more activation at rest compared with task) and an opposing frontoparietal network showing task-related activations (Fox et al., 2005). High demands in tasks that capitalize on controlled, resource-consuming processes have been assumed to leave little available resources for internally guided spontaneous processes, such as mentalizing or mind wandering. From the resting state perspective, the harder the task, the further away from rest, the more diminished DMN involvement. This interpretation supports a resource-consuming account for internally guided thought (Smallwood & Schooler, 2006). On the other hand, McVay and Kane (2010) have claimed that a decreased pool of resources cannot account for a lack of spontaneous internally guided thought, and its underlying DMN activity, supporting their claim with findings showing the opposite: more pronounced episodes of task-unrelated thought during fatigue, in which resources are low to begin with. They also point to findings showing that such mental "lapses" are more evident in patients with either Attention Deficit Hyperactive Disorder (ADHD), or lower working memory span. This account attributes spontaneous associative processes and their underlying neural activity to a failure in the executive control of attention.

Both sides of this debate found either increased default deactivation when attentional requirements increase (leaving little capacity for associative processes), or alternatively decreased deactivation when participants fail to fulfill the attentional requirements of the task at hand. From the perspective of the contextual associations account, the brain is a proactive organ, and as such the brain spontaneously and constantly activates associations. Reductions in DMN activity during attentional tasks thus reflect the reduction in associative activation. This account further suggests that during fatigue or in ADHD, engagement in controlled processes is compromised and internal processing "falls back" to its default tendency, i.e., the proactive associative activations, underlined by heightened default activity. Spontaneous associative processes diminish in the face of additional task requirement as well as take place when these requirements are too demanding to follow. We suggest that the mechanistic questions regarding spontaneity and cost in default activity should shift their focus from "how much" a task demands, to "what" does a task demand. The extent to which one is either able or required to engage in contextual associations is what will predict the magnitude of activation in the contextual associations/default network.

A final issue to consider within the discussion regarding contextual processing and the availability of resources is the idea that the relationship between the two is bi-directional. Extracting contextual associations in itself may play a role in reducing load of various cognitive, perceptual, or emotional processes. Such a premise postulates that the spontaneous unintentional extraction of associations could be a feature rather than a "bug" and is in line with a utilitarian perspective of how our brain operates. The more contextual information is available, the more efficient is the generation of predictions, and the easier is the task.

Ratifying this idea, the contextual associations network has been found to be modulated by expertise. In reviewing the literature on both perceptual and spatial expertise, Cheung and Bar (2012) suggest that as one becomes an expert, and processing becomes holistic, top-down contextual associations are more readily used for predictions and the engagement of the DMN increases. A welcomed "side effect" is that the task is made easier as a product of having contextual associations to generate predictions more readily. In other words, the correlation between reduced task demands and DMN activation could be mediated by the greater accessibility to contextual information. In such cases, it is not decreased task demands that allow spontaneous associative activations, but rather the other way around. The efficient generation of context-based predictions, even if resource-consuming in the short-term, is a long-term economical neural mechanism.

The different forms of task-difficulty and their correlation with activity in the contextual associations network should be experimentally tested apart. This can be done by orthogonally contrasting the level to which one can rely on associative information when performing a task (e.g., mentalizing about the perspective of familiar and unfamiliar others) with other factors that influence the task's difficulty but are irrelevant to associative-based processes (e.g., time constraints). If the availability of cognitive resources is the primary criterion for DMN activity (i.e., the more available resources, the greater DMN activity), one should not find a difference in DMN activation when comparing equally difficult tasks despite possible differences in accessibility to contextual information. Based on our proposal of associative processing in the DMN, however, we expect differences in neural activation within the network to emerge when the compared tasks differ in subject's ability to access contextual information, independent from other factors (e.g., external load). We would predict great activation of the DMN in the case when contextual information is used, while other factors are held constant.

Think Outside the Context Box: Contextual Associations and the Interesting Case of Creativity

One domain that has been extensively linked to default activity in recent years is creativity. Researchers in this field have identified creative thought as involving the use of memory and as relying both on associative and on controlled processes (Beaty et al., 2014, 2016). Divergent thinking, as a proxy for creativity, was suggested to involve DMN in enabling the first stage for creative thought in which remote ideas are activated in an associative manner. Furthermore, it has recently been demonstrated that increased connectivity between the core regions of the DMN and regions of other executive and salience networks is correlated with higher creativity performance, supporting the cooperation of controlled mental navigation processes with spontaneously triggered associative content in creativity (Beaty et al., 2018; Sun et al., 2016). Higher creativity scores in divergent thinking tasks were also found to be correlated with decreased default deactivation during attentionally demanding tasks (Takeuchi et al., 2011). This is predicted if one assumes that divergent thinking relies on associative processing tasks—and therefore activates the DMN.

Another interesting support for the role contextual associations play in creativity comes from a study by Dewhurst, Thorley, Hammond, and Ormerod (2011), which shows that performance in the remote associations task (which requires the generation of numerous associations and is often used to evaluate creative thought; Mednick, 1968) predicts greater susceptibility to the DRM effect (i.e., false recognition of items that were not learned but are contextually associated with a learned list of items). This finding is in line with evidence showing that default activation of contextual information may lead to false memories (Aminoff, Schacter, & Bar, 2008). Taken together, while activating associative information may hamper memory accuracy, it may also facilitate creative thought.

Given the ample evidence regarding the influence of top-down predictions in almost every cognitive function, it is questionable if and how one can really think in a contextually independent manner. The degree to which our minds are contextually restricted may change according to, well, context. As reviewed above, the brain continuously activates associations to generate predictions and prepares itself for whatever comes next. Nevertheless, it is worth noting that some situations benefit from the decreased involvement of these predictions. For example, when meeting someone for the first time we might want to avoid stereotypical or a priori assumptions. Another example is when trying to come up with an original solution to an old problem, and refrain from taking the same unhelpful mental path as before. In such cases, minimizing the reliance on strong associative information may allow access to less predictable judgements and ideas.

Supporting this argument, in a study by Baror and Bar (2016), associative processing in a free associations task was examined under varying conditions of resources availability. This manipulation was aimed to induce varying levels of exploratory behavior by manipulating levels of mental load. It was found that as resources diminished, the diversity in answers decreased, and participants tended to answer with stronger associations, in a more predictable fashion. The authors proposed that when available, resources are invested in inhibition of predictable associations in favor of unique, remote ones. With regard to the debate of whether contextual processing is taxing, under this framework it is reasonable to suggest that even if extracting associative information does not tax resources, applying that information appropriately does. Hence, situations favoring low load, and perhaps such that incorporate physical or mental relaxation such as meditation, may promote open-mindedness and original thought. Accordingly, although an imaging study manipulating free associations under cognitive load has yet to be executed, we hypothesize that changes in associative scope will be underlined by changes in the contextual associations network. This would be in line with the results of Gilmore et al. (2016) regarding imagined vs. experienced events, showing greater DMN activity during remembered events, when associative information is more accessible.

In sum, associations provide the context in which we think (Bar et al., 2007), and therefore the box from which at times we try to think out of. Finding the critical conditions of such a "mental control panel" may be beneficial for situations in which silencing or amplifying the influence of associative processes according to need promotes adaptation, better learning, and creative thought. As we will see next, modulating the level and manner of associative activation is also beneficial in clinical contexts.

Mental Health in Context

Alterations in default activations are found in many clinical diagnoses. Stress, major depressive disorder (MDD), schizophrenia, ADHD, as well as age-related deficits such as Alzheimer's disease and mild cognitive impairment all demonstrate changes in resting state functional connectivity (for review see Greicius, 2008; Broyd et al., 2009). Interestingly, while all these populations show different changes in attentional regulation abilities or other task performance deteriorations, many of them share modifications in associative processes. For example, people suffering from schizophrenia show a pattern of hyper-connectivity within the different nodes of the DMN (Whitfield-Gabrieli et al., 2009). Additionally, among patients experiencing paranoia, selective increased anti-correlations of the DMN with other "task-positive" goal-directed networks are observed (Zhou et al., 2007). These alterations in connectivity within the DMN and between the DMN and other networks seem to correspond with the dysregulated engagement in mental processes, that often comes at the expense of processing external sensory information among patients that are diagnosed with paranoia. By means of contextual associations, this signifies a biased inference process, in which patients disregard perceptual evidence in favor of their predispositions and prior contextual association, to a pathological extent.

The clinical situation in autistic spectrum disorder (ASD) has also been associated with DMN alterations. This network's activity among ASD patients seems to mirror the one observed in schizophrenia, both with regard to the pattern of DMN activity and to the related difficulties in properly relying on context-based predictions. Reduced connectivity within the DMN has been found among ASD patients and is assumed to be related to deficits in self-referential processes (Iacoboni, 2006). In a review by Maras and Bowler (2014), it was suggested that the memory deficit individuals with ASD experience affect their ability to recall personally experienced episodes, integrate information from different domains, retrieve the "gist" of a situation and rely on contextual information. Additionally, when compared with controls, ASD participants show compromised facilitatory effects of emotional cues on memory (Gaigg & Bowler, 2008), implying reduced memory-emotion contextual processing. In a recent review, Van de Cruys et al. (2014) have directly suggested that the core deficit in ASD is to flexibly generate and update associative-based predictions. Furthermore, a recent review focusing on DMN abnormalities in ASD has specifically linked DMN aberrances with the disrupted ability to engage in social information in relation to oneself (Padmanabhan, Lynch, Schaer, & Menon, 2017). We take this claim as evidence of the involvement of DMN activity in drawing appropriate and relevant associations during social interactions in the normative brain.

Another interesting clinical situation that shows alteration in the DMN is ADHD. Studies demonstrate a compromised ability of ADHD patients to coordinate task-unrelated thought with task-related demands, leading to higher distractibility (Gonzalez-Gadea et al., 2015). In a different study, increased distractibility was shown to correlate with decreased default deactivation among children with ADHD (Fassbender et al., 2009). In parallel, adults diagnosed with ADHD demonstrated heightened performance in creativity tasks that require divergent thinking (White & Shah, 2006). Taken together, these findings imply that decreased DMN deactivations are related to increased involvement of associative information in thought, leading to greater distractibility as well as to greater creativity. Furthermore, medications for better attentional control in ADHD have been shown to have negative influence on creativity performance among medicated compared with non-medicated patients (Boot, Nevicka, & Baas, 2017; González-Carpio Hernández & Serrano Selva, 2016). These findings further strengthen the associative account we have proposed and review in this chapter by showing that increased involvement of default activity in associative processes is evident independent of specific task settings, both when associations form an unrelated distraction and when they serve a helpful tool for goal-directed behavior.

The final clinical issue we discuss in this chapter is mood disorders, a mental health problem that is implicated with DMN alterations and is correlated with extensive and dysregulated mind wandering episodes (Berman et al., 2011). Patients with MDD show hyper-connectivity within the DMN, specifically in the subgenual brain region (BA25). The subgenual cortex's increased connectivity with other DMN regions correlates with the duration of depressive episode (Greicius, 2008). Additionally, in patients with mood disorders, the balance between different neural networks is violated and DMN seems to dominate other networks that underlie executive functions. This heightened DMN dominance was found to correlate with increased symptoms of ruminative thought (Hamilton et al., 2011).

Findings from Baror and Bar (2016) address the mood-associations relationship by suggesting that ruminations are akin to load in hindering the ability to generate remote associations. In those studies, cognitive load resulted in narrow and banal associations, and the authors suggest that depression may limit the scope of activated associations in a similar manner, resulting in rumination. The relationship between mood and associations has been proposed (Bar, 2009a, 2009b) and shown in the opposite direction as well, as the progression of thought through broad associative processes was found to improve mood (Brunye et al., 2013; Mason & Bar, 2012).

The brain region that has primarily been suggested to mediate the mood-associativity correlation is the mOFC, a region of the mPFC, which is a critical node in the contextual association network. The mOFC was found to encode independently the associative and the affective values of objects, as well as exhibit correlational patterns between the two (Shenhav et al., 2013). The relationship between associativity and affect is further shown in Trapp et al., (2015) who found that stimuli containing stronger associative information are liked better. It seems that with regard to mood, positive signals may broaden the scope of associations and associative information may promote positive signals.

With the overlap between the DMN and the contextual associations network in mind, one would expect MDD patients to exhibit general deficits in contextual processing, even when the information is not of personal relevance. Recently, it was found that activation in the contextual associations network is altered in non-medicated MDD patients when compared with controls (Harel, Tennyson, Fava, & Bar, 2016). In an fMRI experiment, participants viewed objects that are strongly associated (e.g., a beach chair) or weakly associated (e.g., a bottle) with unique contexts. Reductions in PHC activation were found in depressed patients, compared with healthy controls. These findings contribute to the overarching account linking mood disorders with abnormalities in the activation of contextual associations and the generation of predictions (Bar, 2009b; Barrett, Quigley, & Hamilton, 2016).

In addition, one might consider understanding the default cognitive state of the regions in the DMN by examining what happens when each region is experimentally stimulated. Using electrocorticography to stimulate the ventral-medial temporal region of the brain of an epilepsy patient (Aminoff et al., 2016) this region was shown to process contextual associations. Specifically, when this region was electrically stimulated, possibly akin to what might happen when this region is activated spontaneously during mind wandering, the patient experienced a stream of visual associations retrieved from his long-term episodic memory. This demonstrated that the activation of ventral-medial temporal region resulted in free associative cognitive processing and may reflect what happens when these regions become active during default processing. It also may reflect what may happen if overstimulated in a pathological state, when associations are hyper-activated and may be forced to be tied into the current environment to make sense of the experience, resulting in hallucinations.

To summarize, alterations in resting state connectivity and in the contextual associations network are found among patients with various mental diagnoses. Beyond domain-specific impairments in memory or social processes, many clinical diagnoses include specific impairments in associative processing. Taken together, in the healthy brain, contextual processing takes place by default, and its appropriate execution is often critical for mental well-being.

Concluding Remarks

Continuously, we try to find a unifying framework that would explain the default activity, which underlies thought and behavior in natural as well as in experimental contexts. Such a unifying account should be economic in biological and psychological terms, and intrinsic to both spontaneous and intentional mental processes. As a general framework it should be applicable above and beyond specific forms of content. Furthermore, it is expected to be crucial for maintaining mental health and facilitatory of higher mental functions.

The framework of contextual associations is in line with all these criteria. As elaborated in this chapter, the human brain continuously and proactively generates context-based predictions. Whether generated spontaneously or on demand, these predictions rely on spatial, sensory, emotional, social, and other forms of associations that are stored in memory. Relying on contextual information minimizes demand and is found to facilitate perceptual, social, and creative processes. Contextual predictions are shown to be critically impaired in many clinical situations, and hence essential for maintaining mental health.

Finally, beyond being implicated in all the functions cited above, the neural activation underlying contextual processing overlaps in its brain regions with the DMN and the mentalizing network. Backed both by theory and by evidence, the scheme of contextual processing provides a parsimonious framework, demonstrating how the brain is never truly at rest, always busy with predictions that build on associations, proactive by default.

References

Addis, D. R., Pan, L., Vu, M.-A., Laiser, N., & Schacter, D. L. (2009). Constructive episodic simulation of the future and the past: Distinct subsystems of a core brain network mediate imagining and remembering. *Neuropsychologia, 47*, 2222–2238.

Aminoff, E. M., Gronau, N., & Bar, M. (2007). The parahippocampal cortex mediates spatial and nonspatial associations. *Cerebral Cortex, 17*, 1493–1503.

Aminoff, E. M., Kveraga, K., & Bar, M. (2013). The role of the parahippocampal cortex in cognition. *Trends in Cognitive Sciences, 17*, 379–390.

Aminoff, E. M., Li, Y., Pyles, J. A., Ward, M. J., Richardson, R. M., & Ghuman, A. S. (2016). Associative hallucinations result from stimulating left ventromedial temporal cortex. *Cortex, 83*, 139–144.

Aminoff, E. M., Schacter, D. L., & Bar, M. (2008). The cortical underpinnings of context-based memory distortion. *Journal of Cognitive Neuroscience, 20*, 2226–2237.

Aminoff, E. M., & Tarr, M. J. (2015). Associative processing is inherent in scene perception. *PLoS One, 10*(6), e0128840.

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience, 15*, 670–682.

Andrews-Hanna, J. R. (2012). The brain's default network and its adaptive role in internal mentation. *Neuroscientist, 18*, 251–270.

Andrews-Hanna, J. R., Reidler, J. S., Huang, C., & Buckner, R. L. (2010). Evidence for the default network's role in spontaneous cognition. *Journal of Neurophysiology, 104*, 322–335.

Axelrod, V., Rees, G., Lavidor, M., & Bar, M. (2015). Increasing propensity to mind wander with transcranial direct current stimulation. *Proceedings of the National Academy of Sciences, 112*, 3314–3319.

Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W. Y., Franklin, M. S., & Schooler, J. W. (2012). Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological Science, 23*, 1117. https://doi.org/10.1177/0956797612446024

Baird, B., Smallwood, J., & Schooler, J. W. (2011). Back to the future: Autobiographical planning and the functionality of mind-wandering. *Conscious & Cognition, 20*, 1604–1611.

Bar, M. (2004). Visual objects in context. *Nature Review Neuroscience, 5*, 617–629.

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences, 11*(7), 280–289.

Bar, M. (2009a). The proactive brain: Memory for predictions. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 364*(1521), 1235–1243.

Bar, M. (2009b). A cognitive neuroscience hypothesis of mood and depression. *Trends in Cognitive Sciences, 13*(11), 456–463.

Bar, M., & Aminoff, E. M. (2003). Cortical analysis of visual context. *Neuron, 38*, 347–358.

Bar, M., Aminoff, E. M., & Ishai, A. (2008). Famous faces activate contextual associations in the parahippocampal cortex. *Cerebral Cortex, 18*, 1233–1238.

Bar, M., Aminoff, E. M., Mason, M., & Fenske, M. (2007). The units of thought. *Hippocampus, 17*(6), 420–428.

Bar, M., Aminoff, E. M., & Schacter, D. L. (2008). Scenes unseen: The parahippocampal cortex intrinsically subserves contextual associations, not scenes or places per se. *Journal of Neuroscience, 28*(34), 8539–8544.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmidt, A. M., Dale, A. M., … Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences, 103*, 449–454.

Baror, S., & Bar, M. (2016). Associative activation and its relation to exploration and exploitation in the brain. *Psychological Science, 27*, 776–789.

Barrett, L. F., & Bar, M. (2009). See it with feeling: Affective predictions during object perception. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 364*(1521), 1325–1334.

Barrett, L. F., Quigley, K. S., & Hamilton, P. (2016). An active inference theory of allostasis and interoception in depression. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 371*, 20160011.

Beaty, R. E., Benedek, M., Wilkins, R. W., Jauk, E., Fink, A., Silvia, P. J., & Neubauer, A. C. (2014). Creativity and the default network: A functional connectivity analysis of the creative brain at rest. *Neuropsychologia, 64*, 92–98.

Beaty, R. E., Chen, Q., Christensen, A. P., Qiu, J., Silvia, P. J., & Schacter, D. L. (2018). Brain networks of the imaginative mind: Dynamic functional connectivity of default and cognitive control networks relates to openness to experience. *Human Brain Mapping, 39*, 811–821.

Beaty, R. E., Kaufman, S. B., Benedek, M., Jung, R. E., Kenett, Y. N., Jauk, E., … Silvia, P. J. (2016). Personality and complex brain networks: The role of openness to experience in default network efficiency. *Human Brain Mapping, 779*, 773–779.

Berman, M. G., Peltier, S., Nee, D. E., Kross, E., Deldin, P. J., & Jonides, J. (2011). Depression, rumination and the default network. *Social Cognitive and Affective Neuroscience, 6*, 548–555.

Boot, N., Nevicka, B., & Baas, M., (2017). Creativity in ADHD: Goal-directed motivation and domain specificity. *Journal of Attention Disorders*, 1–10.

Brown, E. C., & Brune, M. (2012). The role of prediction in social neuroscience. *Frontiers in Human Neuroscience, 6*, 147.

Broyd, S. J., Demanuele, C., Debener, S., Helps, S. K., James, C. J., & Sonuga Barke, E. J. (2009). Default-mode brain dysfunction in mental disorders: A systematic review. *Neuroscience Biobehavioral Reviews, 33*, 279–296.

Brunye, T. T., Gagnon, S. A., Paczynski, M., Shenhav, A., Mahoney, C. R., & Taylor, H. A. (2013). Happiness by association: Breadth of free association influences affective states. *Cognition, 127*, 93–98.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function and relevance to disease. *Annals of the New York Academy of Sciences, 1124*, 1–38.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences, 11*, 49–57.

Bar, M. (2004). Visual objects in context. *Nature Review Neuroscience 5*, 617–629.

Cheung, O. S., & Bar, M. (2012). Visual prediction and perceptual expertise. *International Journal of Psychophysiology, 83*, 156–163.

Cheung, O. S., & Bar, M. (2014). The resilience of object predictions: Early recognition across viewpoints and exemplars. *Psychonomic Bulletin & Review, 21*, 682–688.

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences, 106*, 8719–8724.

Dewhurst, S. A., Thorley, C., Hammond, E. R., & Ormerod, T. C. (2011). Convergent, but not divergent, thinking predicts susceptibility to associative memory illusions. *Personality and Individual Differences, 51*, 73–76.

Fassbender, C., Zhang, H., Buzy, W. M., Cortes, C. R., Mizuiri, D., Beckett, L., & Schweitzer, J. B. (2009). A lack of default network suppression is linked to increased distractibility in ADHD. *Brain Research, 1273*, 114–128.

Fenske, M. J., Aminoff, E., Gronau, N., & Bar, M. (2006). Top-down facilitation of visual object recognition: Object-based and context-based contributions. *Progress in Brain Research, 155*, 3–21.

Fox, K. C., Spreng, R. N., Ellamil, M., Andrews-Hanna, J. R., & Christof, K. (2015). The wandering brain: Meta-analysis of functional neuroimaging studies of mind-wandering and related spontaneous thought processes. *NeuroImage, 111*, 611–621.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences, 102*, 9673–9678.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50*(4), 531–534.

Gaigg, S. B., & Bowler, D. M. (2008). Free recall and forgetting of emotionally arousing words in autistic spectrum disorder. *Neuropsychologia, 46*, 2336–2343.

Gilmore, A. W., Nelson, S. M., & McDermott, K. B. (2016). The contextual association network activates more for remembered than for imagined events. *Cerebral Cortex, 26*, 611–617.

González-Carpio Hernández, G., & Serrano Selva, J. P. (2016). Medication and creativity in attention deficit hyperactivity disorder (ADHD). *Psicothema, 28*, 20–25.

Gonzalez-Gadea, M. L., Chennu, S., Bekinschtein, T. A., Rattazzi, A., Beraudi, A., Tripicchio, P., … Ibanez, A. (2015). Predictive coding in autism spectrum disorder and attention deficit hyperactivity disorder. *Journal of Neurophysiology, 114*, 2625–2636.

Greicius, M. (2008). Resting-state functional connectivity in neuropsychiatric disorders. *Current Opinion in Neurology, 21*, 424–430.

Greicius, M. D., Krasnow, B., Reiss, A. L., & Menon, V. (2003). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences, 100*, 253–258.

Hamilton, J. P., Furman, D. J., Chang, C., Thomason, M. E., Dennis, E., & Gotlib, I. H. (2011). Default-mode and task positive network activity in major depressive disorder: Implications for adaptive and maladaptive rumination. *Biological Psychiatry, 70*, 327–333.

Harel, E., Tennyson, R., Fava, M., & Bar, M. (2016). Linking major depression and the neural substrates of associative processing. *Cognitive, Affective, & Behavioral Neuroscience, 16*, 1017–1026.

Hassabis, D., Kumaran, D., & Maguire, E. A. (2007). Using imagination to understand the neural basis of episodic memory. *Journal of Neuroscience, 27*, 14365–14374.

Iacoboni, M. (2006). Failure to deactivate in autism: The co-constitution of self and other. *Trends in Cognitive Sciences, 10*, 431–433.

Kveraga, K., Ghuman, A. S., & Bar, M. (2007). Top-down predictions in the cognitive brain. *Brain Cognition, 65*, 145–168.

Kveraga, K., Ghuman, A. S., Kassam, K. S., Aminoff, E. M., Hamalainen, M. S., Chaumon, M., & Bar, M. (2011). Early onset of neural synchronization in the contextual associations network. *PNAS, 108*(8), 3389–3394.

Maras, K. L., & Bowler, D. M. (2014). Eyewitness testimony in autism spectrum disorder: A review. *Journal of Autism and Developmental Disorders, 44*(11), 2682–2697.

Marchette, S. A., Vass, L. K., Ryan, J., & Epstein, R. A. (2015). Outside looking in: Landmark generalization in the human navigational system. *Journal of Neuroscience, 35*(44), 14896–14908.

Mason, M. F., & Bar, M. (2012). The effect of mental progression on mood. *Journal of Experimental Psychology, 141*(2), 217–221.

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., & Macrae, C. N. (2007). Wandering minds: The default network and stimulus independent thought. *Science, 315*, 393–395.

McVay, J. C., & Kane, M. J. (2010). Does mind wandering reflect executive function or executive failure? Comment on Smallwood and Schooler (2006) and Watkins (2008). *Psychological Bulletin, 136*(2), 188–197, discussion 198–207.

Mednick, S. A. (1968). The remote associates test. *Journal of Creative Behavior, 2*, 213–214.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50*, 655–663.

Mukamel, R., Ekstrom, A. D., Kaplan, J., Iacoboni, M., & Fried, I. (2010). Single-neuron responses in humans during execution and observation of actions. *Current Biology, 20*, 750–756.

Padmanabhan, A., Lynch, C. J., Schaer, M., & Menon, V. (2017). The default mode network in autism. *Biological Psychiatry Cognitive Neuroscience Neuroimaging, 2*, 476–486.

Perry, D., Hendler, T., & Shamay-Tsoory, S. G. (2011). Projecting memories: The role of the hippocampus in emotional mentalizing. *NeuroImage, 54*, 1669–1676.

Raichle, M. E., McLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *PNAS, 98*, 676–682.

Schilbach, L., Eickhoff, S. B., Rotarska-Jagiela, A., Fink, G. R., & Vogeley, K. (2008). Mind at rest? Social cognition as the default mode of cognizing and its putative relationship to the 'default system' of the brain. *Consciousness and Cognition, 17*, 457–467.

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D., & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences, 15*, 319–326.

Sestieri, C., Shulman, G. L., & Corbetta, M. (2017). The contribution of the human posterior parietal cortex to episodic memory. *Nature Reviews Neuroscience, 18*, 183–192.

Shenhav, A., Barrett, L. F., & Bar, M. (2013). Affective value and associative processing share a cortical substrate. *Cognitive, Affective, & Behavioral Neuroscience, 13*(1), 46–59.

Shulman, G., Fiez, J., Corbetta, M., Buckner, R., Miezin, F. M., Raichle, M., & Petersen, S. (1997). Common blood flow changes across visual task: II. Decreases in cerebral cortex. *Journal of Cognitive Neuroscience, 9*, 648–663.

Smallwood, J., & Schooler, J. (2006). The restless mind. *Psychological Bulletin, 132*, 964–958.

Spreng, R. N., & Mar, R. A. (2012). I remember you: A role for memory in social cognition and the functional neuroanatomy of their interaction. *Brain Research, 1428*, 43–50.

Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience, 32*, 489–510.

Sun, J., Chen, Q., Zhang, Q., Li, Y., Li, H., Wei, D., … Qiu, J. (2016). Training your brain to be more creative: Brain functional and structural changes induced by divergent thinking training: The neural plasticity of creativity. *Human Brain Mapping, 37*(10), 3375.

Takeuchi, H., Taki, Y., Hashizume, H., Sassa, Y., Nagase, T., Nouchi, R., & Kawashima, R. (2011). Failing to deactivate: The association between brain activity during a working memory task and creativity. *NeuroImage, 55*, 681–687.

Tamir, D. I., Bricker, A. B., Dodell-Feder, D., & Mitchell, J. P. (2015). Reading fiction and reading minds: The role of simulation in the default network. *Social Cognitive and Affective Neuroscience, 11*, 215–224.

Tamir, D. I., & Mitchell, J. P. (2011). The default network distinguishes construals of proximal versus distal events. *Journal of Cognitive Neuroscience, 23*(10), 2945–2955.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences, 22*(3), 201–212.

Trapp, S., Shenhav, A., Bitzer, S., & Bar, M. (2015). Human preferences are biased towards associative information. *Cognition and Emotion, 29*(6), 1054–1068.

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: Predictive coding in autism. *Psychological Review, 121*, 649–675.

White, H. A., & Shah, P. (2006). Uninhibited imaginations: Creativity in adults with attention-deficit/hyperactivity disorder. *Personality and Individual Differences, 40*, 1121–1131.

Whitfield-Gabrieli, S., Thermenos, H. W., Milanovic, S., Tsuang, M. T., Faraone, S. V., McCarley, R. W., … Seidman, L. J. (2009). Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia. *PNAS, 106*(4), 1279–1284.

Zhou, Y., Liang, M., Tian, L., Wang, K., Hao, Y., Liu, H., … Jiang, T. (2007). Functional disintegration in paranoid schizophrenia using resting-state fMRI. *Schizophrenia Research, 97*, 194–205.

Part V
Mentalizing in Social Interactions and Decision-Making

Computational Approaches to Mentalizing During Observational Learning and Strategic Social Interactions

Caroline J. Charpentier and John P. O'Doherty

In order to navigate our social and connected world, it is key for individuals to be able to learn from other people. Whether it is learning a new skill by observing an expert performing it, learning to seek rewards and to avoid punishments, or making complex strategic decisions, learning from others is prevalent in our daily lives. In this chapter, we give an overview of the behavioral and neural computations at play when we attribute mental states to other agents in order to learn from them. We do so by focusing on two important social learning behaviors and describing the role of mentalizing computations in these processes: observational learning, which involves integrating information received from another agent into one's own beliefs, and strategic thinking, which involves recursive belief inference between agents in order to win a competition or reach a common goal.

The Role of Mentalizing During Observational Learning

The goal of observational learning is to learn which actions and decisions in the environment are good—i.e., likely to lead to positive outcomes—or bad—i.e., likely to lead to negative outcomes—by observing other people performing those actions. Any species endowed with the ability to engage in observational learning has an evolutionary advantage, as it allows individuals to learn about threatening outcomes without having to experience them directly. It can also allow individuals to learn without observing outcomes at all.

Recent research on observational learning in humans has shed light on three possible strategies that can be employed during observational learning (Charpentier & O'Doherty, 2018; Dunne & O'Doherty, 2013). The first strategy is vicarious reward

C. J. Charpentier (✉) · J. P. O'Doherty
Humanities and Social Sciences, California Institute of Technology, Pasadena, CA, USA
e-mail: ccharpen@caltech.edu; jdoherty@caltech.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_24

learning, in which the observer learns by observing another agent making decisions and experiencing outcomes, rather than directly experiencing those outcomes themselves. Similar to experiential learning, the observer is able to form associations between actions and outcomes, but does so through observation. They can then use these learned associations in order to make decisions. A simple computation encoded in the brain to underlie vicarious reward learning is an observational reward prediction error (oRPE), calculated as the difference between the other agent's expected and actual outcome. These oRPEs have been found to be represented in several brain areas, namely vmPFC (Burke, Tobler, Baddeley, & Schultz, 2010; Suzuki et al., 2012), dorsal striatum (Cooper, Dunne, Furey, & O'Doherty, 2012), and ACC (Hill, Boorman, & Fried, 2016). A second observational learning strategy is action imitation, which allows people to learn simply from observing the actions performed by another agent and to repeat or copy the most frequently taken action. Computationally, action imitation learning can be explained in a reinforcement learning framework, with an action prediction error (APE)—the difference between the expected and actual actions of the other agent—reinforcing previously chosen actions positively and unchosen actions negatively. Observers are therefore more likely to perform actions that were also performed by the agent being observed. Neuroimaging results confirmed that APEs are tracked in the brain, specifically in the dmPFC, dlPFC, and inferior parietal lobule (Burke et al., 2010; Suzuki et al., 2012). Finally, the third and more complex observational learning strategy falls under the term "emulation." In emulation learning, observers learn by inferring the other agents' intentions, goals, beliefs, and hidden mental states. The exact computational form of such an inference process is still being investigated, but recent literature suggests that it could take place as a Bayesian inference process, whereby prior beliefs about the other agent's goals are combined with the evidence received from observing the agent's decisions to produce an updated posterior of the inferred beliefs (Charpentier, Iigaya, & O'Doherty, 2020; Charpentier & O'Doherty, 2018; Collette, Pauli, Bossaerts, & O'Doherty, 2017; Devaine, Hollard, & Daunizeau, 2014; Diaconescu et al., 2014). Interestingly, brain areas that have been identified as playing a role in implementing these belief update computations overlap with the mentalizing network: TPJ, pSTS, and dmPFC (Behrens, Hunt, Woolrich, & Rushworth, 2008; Boorman, O'Doherty, Adolphs, & Rangel, 2013; Collette et al., 2017; Charpentier et al., 2020). For example, in Collette et al. (2017), the inference model that best explained participants' behavior was an inverse reinforcement learning (RL) model, whereby instead of learning the value of an action from observing outcomes (classical RL), individuals infer the outcome distribution from observing another agent's actions (inverse RL). The dmPFC was found to contribute to this mechanism by representing the value of the predicted outcomes in agent-referential space (i.e., from the point of view of the agent, not the participant). In addition, the TPJ and pSTS were found to track a learning signal, specifically the entropy or "surprise" predicted by the inverse RL model when observing the other agent's chosen action. While more computationally expensive than vicarious reward learning and action imitation, emulation learning is very adaptive and flexible, can integrate over multiple social signals, and allows the observer to learn from an agent that had different preferences, goals, or even a competing agenda, which will be discussed in more detail in the second part of this chapter.

These strategies can also be used in combination: vicarious reward learning and action imitation (Burke et al., 2010; Suzuki et al., 2012), imitation and emulation (Charpentier et al., 2020), experiential and social learning (Zhang & Gläscher, 2020). Yet, an important outstanding question concerns how it is that people decide or arbitrate between strategies. For example, in a situation where outcomes cannot be directly observed, people can only rely on imitation or emulation in order to learn from observing another agent. The factors that influence the decision to rely on one strategy over the other remain to be elucidated. An interesting hypothesis is that of an arbitration mechanism that would be influenced by the relative estimated reliability of each strategy in the current environment (similar to Lee, Shimojo, & O'Doherty, 2014). In Charpentier et al. (2020), we find that people arbitrate between imitation and emulation learning solely based on the reliability of the emulation system, which depended on the uncertainty of emulation prediction. This reliability signal was represented in the brain, specifically in the ventrolateral prefrontal cortex (vlPFC), but also in the TPJ and ACC. Many factors could play a role in pushing this arbitrator around, such as uncertainty, expertise, or trust in the other agent. For example, if inferring the other agent's intentions becomes more difficult because of increased uncertainty about the evidence provided by observing their behavior, emulation would become more computationally demanding, and it is likely that learning behavior would preferentially rely on imitation. Inversely, if the other agent is deemed incompetent or untrustworthy, simply imitating them may lead to a lot of mistakes and emulation may be favored.

Given this short overview of observational learning, it seems clear that mentalizing plays a role during this process, and does so mainly through emulation learning, which relies on inferring the mental state of another person, whether it is their goals, preferences, beliefs, or intentions. In contrast, vicarious reward learning and action imitation function with simple associative computations, either between action and outcome (vicarious reward learning) or between actions performed by others and actions performed by the self (imitation), suggesting that mentalizing probably doesn't play a role in those strategies. Nonetheless, we note that there may be some degree of overlap in the brain regions involved in the different strategies, for example between imitation and emulation neural signals. Specifically, action prediction errors were found to be encoded in the dmPFC during imitation learning (Burke et al., 2010; Suzuki et al., 2012). The main hypothesis, however, remains that action imitation occurs through the representation of another person's actions in the mirror neuron system, which is active both when an individual performs an action and when they observe another person performing that same action and include regions of the premotor cortex and intraparietal sulcus (Catmur, Walsh, & Heyes, 2009; Lametti & Watkins, 2016; Rizzolatti & Craighero, 2004; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996). This is further supported by a meta-analysis of over 200 fMRI studies (Van Overwalle & Baetens, 2009) comparing the mirror and mentalizing systems and suggesting that the two systems appear to be complementary—rather than one system subserving the other—because they are rarely found to be active together. The recent findings of Charpentier et al. (2020) also confirm this hypothesis, with distinct neural correlates of imitation and emulation update signals, mapping onto the mirror and mentalizing systems, respectively. One situation that may trigger a transition from the mirror to the mentalizing system is when people observing body motions in other people are deliberating about and inferring the goals of these behavioral executions. Additional evidence supports this functional distinction between the two systems during action understanding, with the mirror neuron system suggested to be involved in automatic action identification, perception and understanding of how actions are implemented and the mentalizing system supporting a more controlled representation of why actions are performed by others and understanding the underlying motives and goals (Spunt & Lieberman, 2012, 2013).

The Role of Mentalizing in Strategic Social Interactions

Our ability to learn from observing others can also be applied to cases where there is a mutual and repeated interaction with one or multiple other agents. In everyday social interactions, we don't only rely on others in order to gather information about the world, but we also engage in strategic interactions in which there is an incentive to infer and exploit another person's knowledge or even an incentive to lie or deceive each other in order to maximize our own rewards and outcomes (Lee & Seo, 2016). These behaviors occur in many classic strategy games, such as poker and chess, but also in decisions to cooperate and decisions to engage in prosocial behavior.

An interesting framework to study strategic social interactions in the lab, and to model participants' behavior, is game theory of mind (Camerer, 2003; Yoshida, Dolan, & Friston, 2008). This framework combines predictions of game theory and optimal behavior together with the social component of repeated mutual interactions between agents. It provides a model of how behavior in such social interactions can be optimized through recursive belief inference, specifically making the assumption that "I represent your value function and thoughts, your representation of mine, your representation of my representation of yours, and so on ad infinitum" (Yoshida et al., 2008). In most strategic interactions, an agent's optimal behavior would be to infer their opponent's degree of sophistication—i.e., levels of recursive beliefs inference—and then play using one degree of sophistication higher than their opponent's. An example to illustrate the different degrees of sophistication is the beauty-contest game (Camerer, Ho, & Chong, 2015; Coricelli & Nagel, 2009; Ho, Cambrer, & Weigelt, 1998; Nagel, 1995). In the original version of the game (Keynes, 1936), competitors have to pick the 6 prettiest faces from 100 photographs. The winner is the competitor whose choice is the closest to the average preferences of all competitors. As Keynes pointed out at the time, "it is not a case of choosing those which are really the prettiest, nor even those which average opinion genuinely thinks the prettiest. We have reached the third degree—to anticipating what average opinion expects the average opinion to be." In lab experiments, the game has been formalized as "*p*-Beauty Contests," whereby many players have to choose a number between 0 and 100, and the winner is the person who is closest to the mean of all numbers multiplied by p, with $0 < p < 1$. In the most common setup ($p = 2/3$), a player with no degree of sophistication (level 0) will choose a number randomly, so 50 on average (the expectancy of the uniform distribution of possible answers). A level-1 player will think of other players as being level 0 and will choose 2/3 of 50, so 33 on average. A more sophisticated level-2 player will model others as level-1 and choose 2/3 of 33, so 22 on average. As people progress in their degree of sophistication, they will eventually reach the Nash equilibrium, which would be to choose 0. Studies have shown that most people in the normal population are level-1 or level-2 (Camerer et al., 2015; Ho et al., 1998; Nagel, 1995). Now the question is: does mentalizing play a role in determining the degree of sophistication of an individual who makes inferences about others? Preliminary evidence using the beauty-contest game suggested that it does, as the dmPFC, a region of the mentalizing network, was found to be more active in participants who engage in high relative to low level of inference (Coricelli & Nagel, 2009).

Additional evidence for the role of mentalizing in strategic social interaction has come from studies using a stag-hunt game. In this task, participants interact with another agent and either decide to hunt a rabbit for a small individual payoff or collaborate with the other agent to hunt a stag for a large payoff (Yoshida et al., 2008). Two types of computational models, both implementing recursive belief inference, were fit to the data. One type was a "fixed model," in which the degree of sophistication was assumed to be fixed for each agent throughout the task, and the other type was a "theory of mind model," in which the degree of sophistication is updated after each player's move. The theory of mind model was found to fit participants' data best, suggesting a role for mentalizing when players need to dynamically infer their opponent's strategy and policy. In a subsequent fMRI study, the authors found evidence that signals related to the theory of mind model are represented in the brain (Yoshida, Seymour, Friston, & Dolan, 2010). Specifically, the uncertainty associated with the inference about the other agent's strategy was found in the dmPFC and variations in the other agent's estimated degree of sophistication were associated with activation in the left dlPFC. It is worthwhile to note that this region is outside the classical mentalizing network, and may instead simply reflect the recruitment of executive processes (Chung, Weyandt, & Swentosky, 2014) needed for the increasingly complex inference associated with higher sophistication. The social specificity of this process thus remains open to investigation.

In another set of studies, participants played a competitive game called the "inspector game" in pairs (Hampton, Bossaerts, & O'Doherty, 2008; Hill et al., 2017). This game is a variant of "matching pennies," in which one participant is assigned the role of the employer and the other participant the role of the employee (Fig. 1a). The employer's choices are to inspect or not inspect the employee, while the employee's choices are to work or shirk. The incentives of each player are different, such that the employer has an incentive to not inspect if the employee works or to inspect if the employee shirks, while the employee prefers to shirk if not inspected or work if inspected. Therefore, in order to maximize their outcomes, each participant has to try and predict what the other participant will do next in

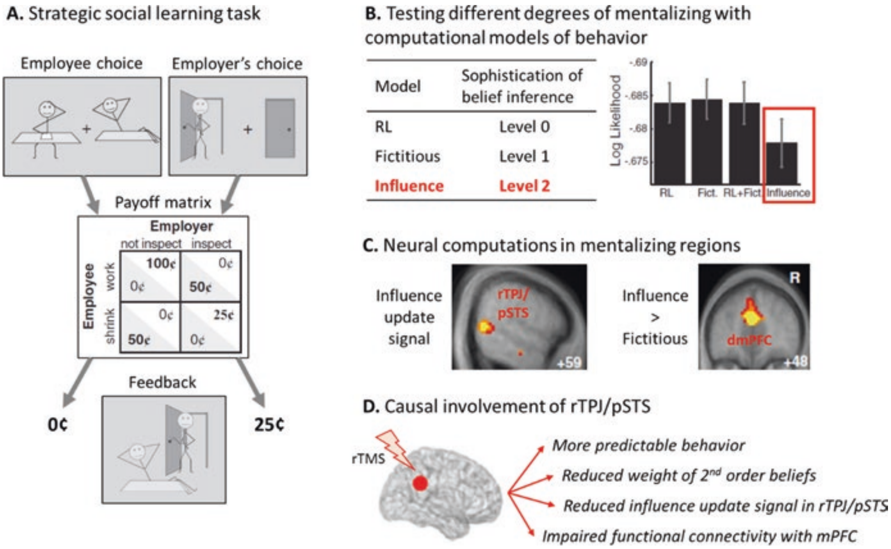

A. Strategic social learning task
Employee choice
Employer's choice
Payoff matrix
Employer
not inspect
inspect
Employee
work
shrink
100¢
0¢
0¢
50¢
0¢
25¢
50¢
0¢
Feedback
0¢
25¢
B. Testing different degrees of mentalizing with computational models of behavior
Model
Sophistication of belief inference
RL
Level 0
Fictitious
Level 1
Influence
Level 2
Log Likelihood
-.69
-.685
-.68
-.675
RL
Fict.
RL+Fict
Influence
C. Neural computations in mentalizing regions
Influence update signal
rTPJ/ pSTS
+59
Influence > Fictitious
R
dmPFC
+48
D. Causal involvement of rTPJ/pSTS
rTMS
More predictable behavior
Reduced weight of 2nd order beliefs
Reduced influence update signal in rTPJ/pSTS
Impaired functional connectivity with mPFC

Fig. 1 A computational social neuroscience approach to study the role of mentalizing in strategic social interactions. A state-of-the-art combination of three methods can be utilized to demonstrate the role of mentalizing in a strategic social learning task. (**a**) In the competitive "inspector game," two players make repeated decisions which have different payoffs depending on the choice of the other player. (**b**) Behaviorally, the role of mentalizing can be tested by comparing different computational models of behavior on the task, showing that the strategy requiring the highest degree of mentalizing outperforms strategies that use some or no mentalizing. (**c**) Neurally, the mentalizing network is recruited and tracks relevant computations predicted by the winning model. (**d**) Finally, manipulating activity in parts of the mentalizing network can show its causal involvement in strategic social interactions. (Adapted from Hampton et al. (2008) and Hill et al. (2017))

order to choose the best action for themselves in consequence. To assess the role of mentalizing in such learning, the authors fitted three types of models to the data, assuming either (1) no mentalizing, (2) an elementary form of mentalizing, or (3) a more sophisticated form of mentalizing (Hampton et al., 2008). The model assuming no mentalizing was a simple reinforcement learning (RL) model, predicting that participants would choose the action that gave the most reward in the recent past. This is equivalent to a level-0 strategy and would be very easy for an opponent to exploit. The model assuming an elementary form of mentalizing employed a strategy called "fictitious play." This strategy learns about the opponent's past actions to predict the upcoming action, thus leading the participant to choose according to that predicted action. Using this strategy involves some representation of the other agent's intentions, similar to a level-1 inference. Finally, the more sophisticated form of mentalizing was called the "influence" learning model, and is equivalent to a level-2 belief inference, in which the participant not only represents their opponent's past actions like in the fictitious play model, but also tracks how their own actions influence the opponent's next play (Fig. 1b). This latter model was found to best explain participants' behavior on the task, confirming a role for mentalizing and second-order representations of others' mental states in this strategic social learning task. Regions of the brain's mentalizing system were also found to track several signals related to the "influence" model. The expected reward associated with the action selected by the participants was encoded in the mPFC at the time of choice, and this signal was better explained by the influence model than by the simpler models, which make different predictions about the expected reward signal. At the time of outcome, activity in the pSTS bilaterally tracked the update in the opponent's inferred strategy as predicted by the influence model, and activity in the dmPFC was associated with the degree to which the influence model outperformed the fictitious play model (Fig. 1c). These fMRI results are therefore consistent with an implication of the mentalizing system, since two key computations were represented in two regions typically involved in mentalizing.

An interesting approach was used in a more recent study to determine whether regions of the mentalizing system are causally involved during strategic social reasoning. In this study, the same inspector game task described above (Hampton et al., 2008) was used in combination with theta-burst repetitive transcranial magnetic stimulation (rTMS) to disrupt neural excitability in the rTPJ and examine whether mentalizing processes are impaired as a result (Hill et al., 2017) (Fig. 1d). Specifically, participants who received rTMS over the rTPJ, compared to a control group who received rTMS over the vertex, were less likely to switch actions and therefore became more predictable for the opponent to exploit. The ability to reason about the influence of the player's own actions on the opponent's response (second-order beliefs) was also found to be significantly reduced in the participants whose rTPJ activity was disrupted. The influence update signal in the rTPJ/pSTS was also reduced by the stimulation, suggesting that disrupting neural excitability in this region impaired its ability to efficiently encode the necessary social learning signal. Interestingly, the authors also examined long-range effects of rTPJ stimulation by examining neural activity in the dmPFC and vmPFC, as well as functional connectivity between the rTPJ and these regions. Replicating previous findings from Hampton et al. (2008), individual differences in the dmPFC influence update signal were found to predict how likely participants were to rely on the influence over fictitious strategy. This relationship was not affected by rTPJ stimulation, suggesting that the representation of this second-order influence model in the dmPFC does not exclusively depend on inputs from the rTPJ. However, functional connectivity between the rTPJ and frontal regions was found to be affected by the stimulation in two ways: (1) reduced functional connectivity between rTPJ and a more dorsal region of the dmPFC at the time of feedback (relative to baseline) and (2) reduced modulation of functional connectivity between rTPJ and vmPFC by the influence update signal. In summary, this study provided crucial evidence to support a causal role for rTPJ in both the behavioral and neural computations associated with mentalizing; in other words, demonstrating that the mentalizing system is necessary for people to be able to learn how their own actions influence their opponent's future behavior.

Mentalizing regions were also found to play a role in a slightly different type of strategic social interaction—advice giving (Hertz et al., 2017). In a task aimed at eliciting this phenomenon, the participant plays the role of one of two advisers who give advice to a client and have to compete for social influence in order for the client to choose them over the other adviser. Activity in the rTPJ was found to represent whether the participant was chosen by the client or not, which, according to the model, played a role in subsequent strategic influence over the client. Activity in the mPFC encoded relative merit, or advice accuracy relative to the other adviser. In a multi-round economic exchange game (Xiang, Ray, Lohrenz, Dayan, & Montague, 2012), one player is an investor deciding which fraction of a $20 endowment to share with a trustee, the fraction is tripled, and the trustee decides which fraction of that triple amount to repay to the investor. Computational modelling of behavior allowed classifying each investor, who also underwent fMRI, into level-0 (about 50% of investors), level-1 (about 25%) or level-2 (about 25%) players. Different patterns of neural activations were found in the three groups; specifically, the rTPJ was found to track first-order interpersonal prediction errors (when repayments were revealed) more strongly in level-2 compared to level-0 players.

While all the work describe above has focused on human mentalizing and social learning, there is also limited—but nonetheless interesting—evidence that non-human primates can engage in complex strategic social interactions. Two recent studies show that monkeys (1) can predict their opponent's actions and counter a possible exploitation by the opponent (Seo, Cai, Donahue, & Lee, 2014), and (2) can recursively infer another agent's intentions to decide whether to cooperate or not (Ong, Madlon-Kay, & Platt, 2018).

In the first study (Seo et al., 2014), the authors recorded from monkeys' dmPFC neurons while the animals performed a biased matching pennies game against a computer opponent. In the game both players choose between two targets. If they choose the same target, the monkey wins a point; if they choose different targets, the animal either loses a point (risky option) or gets nothing (safe option). The computer opponent's behavior was such that if the monkey chose the risky or safe option more frequently than predicted by the optimal strategy, this behavior was exploited by the computer. Therefore, the monkey has an incentive to not be too predictable. This is exactly what the behavioral model showed. Contrary to the predictions of simple reinforcement learning, the animals' actions did not only depend on their previous outcomes, but also on their previous actions, suggesting that they learn to change their action patterns in order to not be exploited by the computer. In addition, when the computer's actions were predictable, the monkeys were able to exploit them to maximize their payoffs. Neurons in the dmPFC were found to represent the integration of both past outcomes and past choices, as predicted by higher-order inference about the opponent. Stay versus switch choices were decoded from dmPFC activity, such that the difference in decoding accuracy for switch versus stay choices was predictive of the extent to which the monkey's switch choices deviated from the simple reinforcement learning algorithm. In other words, the more decisions to switch were consistent with strategic thinking, the more dmPFC neurons' activity could decode those decisions.

In another study, Ong and colleagues sought to provide evidence for TPJ homolog regions in non-human primates and to test its role in strategic interactions (Ong et al., 2018). To do so, they recorded from middle STS (mSTS) neurons while monkeys played a version of the "chicken" game. The game is somewhat similar to the stag-hunt game described above (Yoshida et al., 2008, 2010). In this game, two monkeys are facing each other and moving a joystick to either go straight or yield to the side. If they both go straight, they will "crash" into each other and receive no reward. If they both yield, they will get a medium cooperation reward. If one monkey yields and the other goes straight, the monkey who yields gets a small reward and the monkey who goes straight gets a large reward. This task allowed testing whether the monkeys would rather coordinate in order to obtain a cooperative reward, or rather compete to pursue an individual reward at the expense of their opponent. Interestingly, the payoffs varied across trials such that a mixed strategy switching between cooperating and competing was optimal. Behavioral results showed that monkeys largely avoided going straight and crashing, suggesting they relied on the other player's behavior to also guide their choice. Specifically, computational models of behavior were tested with different degrees of sophistication. The best-fitting model was found to be the one with the most sophistication, including both a representation of the other monkey's maximum payoffs and learning about the other monkey's strategy via a strategy prediction error (SPE), suggesting an engagement of mentalizing function. Interestingly, mSTS neurons were found to selectively respond to reward obtained cooperatively, but not to rewards obtained selfishly. Some neurons in both mSTS and ACC were also found to encode the opponent's strategy, as predicted by the model. Overall, this very promising line of work suggests that non-human primates also engage in some form of mentalizing, which relies on similar brain networks as humans, to learn from another agent in the context of strategic interactions.

Those two studies provide evidence for a role of mentalizing in social interaction in macaque monkeys. In another study (Devaine et al., 2017), the authors were able to compare mentalizing abilities from seven non-human primate species—specifically lemurs, macaques, mangabeys, orangutans, gorillas, and chimpanzees—to test whether mentalizing abilities and degree of sophistication are better explained by social network complexity (as indexed by group size) or by cognitive capacity (as indexed by brain volume). All animals from the 7 species (39 in total) played simple dyadic games against artificial players with different degrees of sophistication. Using computational models of behavior, mentalizing abilities on these games were found to be more strongly associated with brain volume rather than social network complexity, suggesting that mentalizing abilities seem to be limited by neurobiological factors and overall cognitive capacity. In addition, comparing the animals' performance with human players, the authors also conclude that great apes' mentalizing abilities still fall short of that of humans.

Conclusions and Future Directions

In this chapter, we have explored evidence suggesting that the brain's mentalizing network—dmPFC, pSTS, and TPJ—is involved not only in learning from another agent by inferring their intentions, goals, and beliefs, but can also perform complex computations of mental state inference during strategic social interactions. We highlight that computational models of belief inference, model-based neural activations in the mentalizing network, and causal manipulation of these neural computations (Fig. 1b–d) constitute three valuable methods for examining the role of mentalizing in observational learning and strategic social thinking, especially when used in combination. In this neuro-computational approach, specific mathematical variables predicting behavior are extracted from a computational model and can be directly regressed against brain activity, thus refining our understanding of how exactly a particular process is implemented in the brain. This approach, across the many studies described in this chapter, has provided us with a novel perspective about how different areas of the mentalizing network represent specific mentalizing computations.

Overall, more studies are needed to provide a more integrated account of the computational mechanisms associated with observational and strategic social learning, both at the behavioral and neural level. We would like to highlight some open questions that have yet to be addressed:

- Is mentalizing required for social learning? Recent evidence suggests that by disrupting activity in the brain's mentalizing network (Hill et al., 2017), as noted, or by studying a clinical population with disrupted mentalizing ability (Rosenthal, Hutcherson, Adolphs, & Stanley, 2019), we can show that mentalizing is necessary for some particular social learning processes. However, this evidence is still extremely limited and preliminary and more studies are needed to generalize these finding to a range of observational learning and strategic social interaction tasks.
- How specific are the computations associated with a particular social learning strategy? In many studies using a computational modelling approach to behavior, one "winning" model and the computations predicted by this model are selected because of their greater explanatory power. However, the specificity of these computations is rarely tested and it is possible that several models would result in the same behavioral and neural computations, thus questioning the specificity of the particular "winning" model.
- Is mentalizing involved in the arbitration between two social learning strategies? When decisions in a social learning task are found to be a combination of two strategies, it is unclear whether mentalizing abilities, and the mentalizing network, play a role in arbitrating between these strategies. For example, is mentalizing required to decide between relying on imitation versus emulation when learning from another agent? Or to decide between level-1 and level-2 reasoning during strategic interactions?
- How do the different brain regions involved in these processes functionally interact? Very few studies to date have examined functional connectivity between regions of the mentalizing network in the context of social learning computations. Preliminary evidence (Hill et al., 2017; Zhang & Gläscher, 2020) suggests that connectivity between the TPJ and the prefrontal cortex would be a good candidate to investigate further.

- Does mentalizing play a different role when the goal of social learning is to obtain rewards versus avoid threats? Most studies covered in this chapter examine observational or strategic learning tasks in which the participant's goal is usually to maximize some positive outcomes (e.g., monetary rewards). However, investigations of social learning to avoid threat are much less common (for an example, see Parnamets, Espinosa, & Olsson, 2020), and it is unknown whether and how behavioral and neural computations would differ between positive and negative contexts.

References

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature, 456*(7219), 24524–24529. https://doi.org/10.1038/nature07538

Boorman, E. D., O'Doherty, J. P., Adolphs, R., & Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron, 80*(6), 1558–1571. https://doi.org/10.1016/j.neuron.2013.10.024

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences, 107*(32), 14431–14436. https://doi.org/10.1073/pnas.1003111107

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Camerer, C. F., Ho, T. H., & Chong, J. K. (2015). A psychological approach to strategic thinking in games. *Current Opinion in Behavioral Sciences, 3*(1), 157–162. https://doi.org/10.1016/j.cobeha.2015.04.005

Catmur, C., Walsh, V., & Heyes, C. (2009). Associative sequence learning: The role of experience in the development of imitation and the mirror system. *Philosophical Transactions of the Royal Society: Biological Sciences, 364*(1528), 2369–2380. https://doi.org/10.1098/rstb.2009.0048

Charpentier, C. J., & O'Doherty, J. P. (2018). The application of computational models to social neuroscience: Promises and pitfalls. *Social Neuroscience, 13*(6), 637–647. https://doi.org/10.1080/17470919.2018.1518834

Chung, H. J., Weyandt, L. L., & Swentosky, A. (2014). The physiology of executive functioning. In S. Goldstein & J. Naglieri (Eds.), *Handbook of executive functioning* (pp. 13–27). New York, NY: Springer.

Collette, S., Pauli, W. M., Bossaerts, P., & O'Doherty, J. P. (2017). Neural computations underlying inverse reinforcement learning in the human brain. *eLife, 6*, e29718. https://doi.org/10.7554/eLife.29718

Cooper, J. C., Dunne, S., Furey, T., & O'Doherty, J. P. (2012). Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. *Journal of Cognitive Neuroscience, 24*(1), 106–118. https://doi.org/10.1162/jocn_a_00114

Coricelli, G., & Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences, 106*(23), 9163–9168. https://doi.org/10.1073/pnas.0807721106

Charpentier, C. J., Iigaya, K., & O'Doherty, J. P. (2020). A neuro-computational account of arbitration between choice imitation and goal emulation during human observational learning. *Neuron*, *106*(4), 687–699. https://doi.org/10.1016/j.neuron.2020.02.028

Devaine, M., Hollard, G., & Daunizeau, J. (2014). The social Bayesian brain: Does mentalizing make a difference when we learn? *PLoS Computational Biology, 10*(12), e1003992. https://doi.org/10.1371/journal.pcbi.1003992

Devaine, M., San-Galli, A., Trapanese, C., Bardino, G., Hano, C., Saint Jalme, M., … Daunizeau, J. (2017). Reading wild minds: A computational assay of theory of mind sophistication across seven primate species. *PLoS Computational Biology, 13*(11), e1005833. https://doi.org/10.1371/journal.pcbi.1005833

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., … Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Computational Biology, 10*(9), e1003810. https://doi.org/10.1371/journal.pcbi.1003810

Dunne, S., & O'Doherty, J. P. (2013). Insights from the application of computational neuroimaging to social neuroscience. *Current Opinion in Neurobiology, 23*(3), 387–392. https://doi.org/10.1016/j.conb.2013.02.007

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences, 105*(18), 6741–6746. https://doi.org/10.1073/pnas.0711099105

Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C., & Bahrami, B. (2017). Neural computations underpinning the strategic management of influence in advice giving. *BioRxiv*, 121947. https://doi.org/10.1101/121947

Hill, C. A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J. P., & Ruff, C. C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience, 20*, 1142–1149. https://doi.org/10.1038/nn.4602

Hill, M. R., Boorman, E. D., & Fried, I. (2016). Observational learning computations in neurons of the human anterior cingulate cortex. *Nature Communications, 7*, 12722. https://doi.org/10.1038/ncomms12722

Ho, T. H., Cambrer, C., & Weigelt, K. (1998). Iterated dominance and iterated best response in experimental "p-beauty contests". *American Economic Review, 88*(4), 947–969.

Keynes, J. M. (1936). *The general theory of employment, interest and money*. New York, NY: Harcourt Brace andCompany.

Lametti, D. R., & Watkins, K. E. (2016). Cognitive neuroscience: The neural basis of motor learning by observing. *Current Biology, 26*(7), R288–R290. https://doi.org/10.1016/j.cub.2016.02.045

Lee, D., & Seo, H. (2016). Neural basis of strategic decision making. *Trends in Neurosciences, 39*(1), 40–48. https://doi.org/10.1016/j.tins.2015.11.002

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron, 81*(3), 687–699. https://doi.org/10.1016/j.neuron.2013.11.028

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review, 85*(5), 1313–1326. http://www.aeaweb.org/aer/.

Ong, W. S., Madlon-Kay, S., & Platt, M. L. (2018). Neuronal mechanisms of strategic cooperation. *BioRxiv*, 500850. https://doi.org/10.1101/500850

Pärnamets, P., Espinosa, L., & Olsson, A. (2020). Physiological synchrony predicts observational threat learning in humans. *Proc. R. Soc. B.,* 287, 20192779. http://doi.org/10.1098/rspb.2019.2779

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*(1), 169–192. https://doi.org/10.1146/annurev.neuro.27.070203.144230

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*(2), 131–141. https://doi.org/10.1016/0926-6410(95)00038-0

Rosenthal, I. A., Hutcherson, C. A., Adolphs, R., & Stanley, D. A. (2019). Deconstructing theory-of-mind impairment in high-functioning adults with autism. *Current Biology, 29*(3), 513–519. https://doi.org/10.1016/j.cub.2018.12.039

Seo, H., Cai, X., Donahue, C. H., & Lee, D. (2014). Neural correlates of strategic reasoning during competitive games. *Science, 346*(6207), 340–343. https://doi.org/10.1126/science.1256254

Spunt, R. P., & Lieberman, M. D. (2012). Dissociating modality-specific and supramodal neural systems for action understanding. *Journal of Neuroscience, 32*(10), 3575–3583. https://doi.org/10.1523/JNEUROSCI.5715-11.2012

Spunt, R. P., & Lieberman, M. D. (2013). The busy social brain: Evidence for automaticity and control in the neural systems supporting social cognition and action understanding. *Psychological Science, 24*(1), 80–86. https://doi.org/10.1177/0956797612450884

Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., … Nakahara, H. (2012). Learning to simulate others' decisions. *Neuron, 74*(6), 1125–1137. https://doi.org/10.1016/j.neuron.2012.04.030

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage, 48*(3), 564–584. https://doi.org/10.1016/j.neuroimage.2009.06.009

Xiang, T., Ray, D., Lohrenz, T., Dayan, P., & Montague, P. R. (2012). Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS Computational Biology, 8*(12), e1002841. https://doi.org/10.1371/journal.pcbi.1002841

Yoshida, W., Dolan, R. J., & Friston, K. J. (2008). Game theory of mind. *PLoS Computational Biology, 4*(12), e1000254. https://doi.org/10.1371/journal.pcbi.1000254

Yoshida, W., Seymour, B., Friston, K. J., & Dolan, R. J. (2010). Neural mechanisms of belief inference during cooperative games. *The Journal of Neuroscience, 30*(32), 10744–10751. https://doi.org/10.1523/JNEUROSCI.5895-09.2010

Zhang, L., & Gläscher, J. (2020). A brain network supporting social influences in human decision-making. *Science Advances*, *6*(34), eabb4159. https://doi.org/10.1126/sciadv.abb4159

Mentalizing in Value-Based Social Decision-Making: Shaping Expectations and Social Norms

Claudia Civai and Alan Sanfey

Value-Based Decision-Making

Everyday life is defined by choice: deciding what to have for breakfast, what clothes to wear, what career to pursue, or who to befriend. All complex cognitive processes involve evaluating alternatives and eventually selecting one option that is deemed preferable. Understanding how value is assigned to each alternative is central to explaining the psychological mechanisms of decision-making. One recently proposed, and effective, means to investigate value-based choices is a neuroeconomic approach, which integrates strengths of different disciplines, providing a clear operationalization of value using utility functions (from economics) and fine-grained and multi-layered explanations of the cognitive and neural mechanisms involved in the evaluation process (from psychology and neuroscience) (Rangel, Camerer, & Montague, 2008). From this rich body of literature, the existence of a neural circuit underpinning value computation has emerged, encompassing subcortical and cortical areas; a recent meta-analysis (Clithero & Rangel, 2014) found that the posterior cingulate cortex (PCC), the ventral striatum (VStr), and the medial part of the orbitofrontal, also referred to ventromedial prefrontal cortex (vmPFC)[1] play crucial roles in integrating external and internal information, and

[1] Dixon, Thiruchselvam, Todd, and Christoff (2017), in their recent exhaustive review on the anatomical and functional parcellation of the prefrontal cortex, distinguish a lateral OFC, underpinning the evaluation of external stimuli in relation to internal goals, and a medial OFC, or ventromedial prefrontal cortex (vmPFC), integrating the evaluation of the external stimuli with internally generated scenarios and contributing to value-based decision-making. In relation to

C. Civai (✉)
Division of Psychology, School of Applied Sciences, London South Bank University,
London, UK
e-mail: civaic@lsbu.ac.uk

A. Sanfey
Donders Institute for Brain, Cognition and Behaviour, Radboud University,
Nijmegen, The Netherlands

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_25

thus eventually determining the subjective value of a choice option. This system underpins the evaluation of any stimulus characterized by relevant valence in the context of the decision process, with the system commonly activated for primary and secondary incentives, such as food or money (Knutson, Westdorp, Kaiser, & Hommer, 2000; McClure, Ericson, Laibson, Loewenstein, & Cohen, 2007), as well as social incentives, such as praise or good reputation (Izuma, Saito, & Sadato, 2008; Rilling et al., 2002).

Whether social decision-making, e.g. to decide to cooperate with another, as opposed to individual decision-making, e.g. to decide what to eat for dinner, should be considered as a separate aspect of the cognitive system, or whether the same basic processes apply to both individual and social decisions, is an issue of debate; these two accounts are, however, not necessarily mutually exclusive (see Ruff & Fehr, 2014). On the one hand, there are cognitive mechanisms whose existence are potentially uniquely associated with socio-cognitive processing, such as Theory of Mind (ToM) and empathy, in that their involvement in cognitive processing depends on the presence of other individuals. Supporting the idea that social cognition is a unique mechanism, there is evidence of brain areas, such as the temporo-parietal junction (TPJ), that are specifically associated with the multidimensional domain of ToM (see Schurz, Radua, Aichhorn, Richlan, & Perner, 2014 for a meta-analysis; Lee & Seo, 2016). On the other hand, when considering incentives and motivation, the neuroscientific findings are more in line with the idea of a common mechanism that evaluates both social and non-social rewards and guides both types of behaviour accordingly (Clithero & Rangel, 2014; Fehr & Camerer, 2007; Ruff & Fehr, 2014).

Economic theory, in particular behavioural economics, has formalized the social aspect of these processes by incorporating the other agent into the utility function, which describes mathematically the value attached to a decision outcome. Theories of other-regarding preferences are so-called because they consider the presence of other individuals and their status (e.g. economic payoff) as an integral part of the scenario that a person evaluates when making a social decision. These models incorporate the other's payoff as a parameter in summarizing the final outcome of the decision: for example, the parameter value representing the payoff to the other agent is crucial in determining whether, and how much, an individual is averse to inequality (Fehr & Schmidt, 1999); similarly, assessing the importance ascribed to the other's intentions is central to understanding decisions to reciprocate good, as well as bad, behaviour (Charness & Rabin, 2002; Dufwenberg & Kirchsteiger, 2004). The utility associated with these complex decision outputs is associated with neural signals in the vmPFC and striatum, key areas for encoding both social and non-social rewards. Tricomi and colleagues, for instance, found that activation in the ventral striatum was stronger when people could increase the payoff of a disadvantaged group, at a cost to themselves, and re-establish equality in the exchange (Tricomi, Rangel, Camerer, & O'Doherty, 2010); similarly, the striatum showed increased activation when people were given the chance to engage in costly punishment of injustice and unfairness (de Quervain et al., 2004; Stallen et al., 2018; Strobel et al., 2011).

mentalizing, while the OFC evaluates others on the basis of external features, more dorsal areas of the medial prefrontal cortex are involved in evaluating others' mental states.

Mentalizing in Strategic and Prosocial Value-Based Interactions

How is the other-regarding element integrated into the valuation process? Computing social signals and integrating them into subjective valuation involves mentalizing: choosing to set aside one's self-interest to be generous and charitable, or simply fair, requires the ability to take the other's perspective into consideration, and to understand the feelings and beliefs of the other. As previously mentioned, these cognitive mechanisms, which are intrinsically linked to the social context, are underpinned by specific brain areas, encompassing both posterior, i.e. temporoparietal junction (TPJ) and PCC, and anterior, i.e. medial prefrontal cortex (mPFC), regions (Schurz et al., 2014).

These mentalizing areas have been implicated in different aspects of social decision-making tasks, and the evaluation of other-regarding preferences in particular. For example, the activity of mPFC, and especially its dorsal region, has been positively linked to understanding and correctly predicting others' preferences (Kang, Lee, Sul, & Kim, 2013), as well as other-regarding values in a reward-based task, i.e. the utility that others would derive from a specific choice made by the decision maker themselves (Sul et al., 2015). Interestingly, Sul and colleagues found a gradient in the mPFC, whereby dorsomedial areas (dmPFC) represented other-regarding values and ventromedial areas (vmPFC) correlated with self-regarding values. Selfish individuals showed a clear regional differentiation, with vmPFC active for self-regarding values and dmPFC active for other-regarding values; conversely, prosocial individuals, while showing a higher vmPFC for personal rewards, lacked this gradient for other-regarding values, instead demonstrating equal strength of activation in both regions. This may be in line with the hypothesis that vmPFC computes an overall value signal after integrating different pieces of information (e.g. Roy, Shohamy, & Wager, 2012), which, in the case of other-regarding values, is higher for prosocial as compared to selfish individuals. This interpretation would concur with other findings, such as those by Suzuki et al. (2012), who showed that vmPFC encodes the shared representation of self and other reward-prediction error, defined as the difference between what one gets and what one expected to get; and by Hutcherson and colleagues, who identified vmPFC (Bechara, 2005) as the region that encoded both self and other rewards in a simple dictator game (see Box 1), where participants can decide how to split a sum of money between themselves and another powerless player (Hutcherson, Bushong, & Rangel, 2015). Conversely, dmPFC may be an area that is specifically recruited to compute other-regarding values.

Box 1 Experimental Tasks

Adapted from Vavra, van Baar, and Sanfey (2017); for a review of the use of game theory and its paradigms in neuroeconomics, see Civai and Hawes (2016).

Behavioural economics and game theory offer a wide range of structured paradigms that can easily be adapted to a laboratory-based exploration of decision-making, usually involving a multi-player structure where participants are asked to make decisions. Some of the games that have been used in the context of investigation of social norms perception and compliance, and that are referred to in the current chapter, are described below.

The ultimatum game (UG; Güth, Schmittberger, & Schwarze, 1982) is a game where two players decide, sequentially, how to split a sum of money. The first player (proposer) is given a sum of money, e.g. $10, and has to decide how to divide this money with the second player (responder). The proposer makes offers to the responder, who has to either accept or reject these offers: if they accept, the money is divided as the proposer decided; if they reject, none of the players gets anything. Importantly, the game is often anonymous and played as a single round (one-shot game), therefore there is no room for negotiation. Some players may be considered strictly egalitarian (proposers always offering half of the share; responders always rejecting less than half of the share) or strictly rational (proposers always offering the smallest unit; responders always accepting any offer larger than 0). However, experimental evidence has been consistently showing that proposers tend to offer a fair share, and responders reject unfair offers (Camerer, 2003). This highlights the role played by mentalizing: the proposer will offer the smallest amount they believe will be accepted by the responder; on the other hand, the responder will accept any offer that is deemed fair, considering the circumstances and the proposer's intentions (Falk, Fehr, & Fischbacher, 2003).

The dictator game (DG) is very similar to the UG; the only difference is that the responder is powerless and does not have the chance to reject the offer. As a consequence, the first player (dictator) decides the allocation of the monetary sum and their decision does not have any consequence within the game. The strategic motivation for being fair is now removed, and therefore any monetary transfer in this game can be considered as genuine generosity; in this case, mentalizing and particularly empathic concerns may explain the altruistic behaviour (Edele, Dziobek, & Keller, 2013).

Third party games are often adapted versions of the UG or DG, where one additional player plays the role of the observer. In these games, the observer is required to decide whether or not to react to an injustice, by spending their own resources, when their payoff had not been affected by the injustice. The observer may react by punishing the perpetrator (e.g. an unfair dictator; Fehr & Fischbacher, 2004; Strobel et al., 2011) or by compensating the victim (Stallen et al., 2018). Mentalizing and other-regarding concerns are involved in these decisions, in particular when choosing to compensate the victim (Civai, Huijsmans, & Sanfey, 2019; Leliveld, van Dijk, & van Beest, 2012).

The trust game (TG; Berg, Dickhaut, & McCabe, 1995) is a two-player game widely employed to investigate trust and reciprocity. One player (investor) is endowed with a sum of money and can decide how much to transfer to the second player (trustee). The rules of the game establish that any amount transferred is multiplied by a fixed factor, e.g. four. For example, if the investor transfers $5, the trustee would receive $20; at this point, the trustee decides how much of this amount to transfer back to the investor, if any. Because trustee can decide to transfer nothing back, the decision of the investor to transfer something can be interpreted as trust in the second player. On the other hand, the trustee's decision to return any amount is interpreted as reciprocity. Similar to the UG, the investor will use mentalizing abilities to predict the trustee's willingness to reciprocate; in turn, the trustee may use the ability to predict the investor's mental state to decide how much to transfer back: this is true in particular for the guilt-free trustees who reciprocate to match the investor's expectations, as explained in the main text (van Baar, Chang, & Sanfey, 2019).

The prisoner's dilemma, first formalized in the 1950s by Flood and Dresher, is a game in which two players must decide whether to cooperate with each other or defect, knowing that cooperation would lead to the maximum outcome for both players. The game can be played simultaneously or sequentially, eliciting a tit-for-tat strategy; as shown by neuroimaging studies, mentalizing is one of the core mechanisms to guide the decision (Rilling et al., 2004).

As previously mentioned, all these games are often anonymous and one-shot. In neuroimaging studies, since it is necessary to have multiple observations, the so-called single-shot multi-round games are employed: each participant plays the game multiple times, on each round paired with a new partner. Repeated paradigms, i.e. having participants interacting with the same player more than once, are employed when the focus of the investigation is learning process. As in Heijne and Sanfey (2015) for example, interacting with the same partner more than once allows participants to learn whether to stay or leave the relationship. In conclusion, game theory offered a set of structured and flexible paradigms well suited for investigating social interaction in a laboratory context.

There is ample evidence that supports the involvement of TPJ, specifically in the right hemisphere, in considering other-regarding preferences. For example, Hutcherson et al's neurocomputational model found that the activation of this area positively correlated with the amount of money allocated to the other person in the dictator game, suggesting that an other-regarding value signal is already encoded here. Morishima and colleagues performed a voxel-based morphometry analysis and found that grey matter volume of the right TPJ was positively associated with people's altruistic preferences in advantageous inequality situations, i.e. when participants were better off than their task partner (Morishima, Schunk, Bruhin, Ruff, & Fehr, 2012). Other findings support the link between value encoding and mentalizing, suggesting that the subjective evaluation of social stimuli depends on the strength of the functional connectivity between subjective-value areas (vmPFC) and social cognition areas (TPJ) (Smith, Clithero, Boltuck, & Huettel, 2014). For example, Strombach and colleagues found that the connectivity between these two areas was greater when people chose a generous rather than a selfish option in a social decision-making task, suggesting the integration of social signals into the final subjective evaluation in order to guide decisions (Strombach et al., 2015).

Mentalizing Shapes Expectations

Why does the other's perspective need to be integrated into the subjective valuation that eventually determines decisions in social contexts? The goal of the valuation process is to determine the optimal outcome for the decision maker; hence, being able to predict the various outcomes of all potential choice options is fundamental in order to select the best one. By making the other's state of mind available to the decision maker, mentalizing allows for the prediction of the other's behavioural reactions in different scenarios, adding a crucial element to the choice process. Let us consider, for example, the case of the ultimatum game (See Box 1; Güth et al., 1982): in this task, the first player (proposer) is asked to divide a sum of money, for example $10, with the second player (responder), who can decide whether to accept the offer of the proposer or reject it. If the offer is rejected, both players end up with zero, with no possibility to reopen negotiation. The Nash equilibrium for this game predicts that the proposer will offer the smallest amount of money that the responder is willing to accept; if the proposer thinks that the responder is an economically rational player and will accept anything higher than zero, then they will offer the minimum, e.g. $1, or even less. On the other hand, if the proposer thinks that the responder is strictly egalitarian, they will offer half of the sum, because they fear that anything less will be rejected. Whichever solution is chosen by the proposer, it is clear that their allocation decision is based on the proposer's evaluation of the responder's perspective, and the prediction of how this will subsequently drive their behaviour to either accept or reject. In more general terms, it is possible to say that decisions are driven by the expectations that we hold about the outcome of a certain social interaction (e.g., keeping money because the responder has accepted our ultimatum game's offer), and these expectations are in turn shaped by our ability to take the other's perspective and predict their behaviour.

Predicting outcomes is therefore vitally important in order to select the right strategy, where, the 'right strategy' is one that delivers the preferred outcome in that context (e.g. self-interest, altruistic/prosocial, egalitarian). A useful example to clarify this concept comes from developmental science. Mentalizing ability develops with age, and younger children are generally less generous than older children (Benenson, Pascoe, & Radmore, 2007). If taking someone else's perspective automatically resulted in greater generosity, then we should expect generosity to increase with perspective taking and mentalizing abilities. However, the results of a study by Cowell and colleagues challenge this position: in fact, when 3–5-year-olds were asked to play as proposers, or dictators, in a dictator game, which is similar to the ultimatum game with the crucial exception that responders passively receive offers without the opportunity to change the outcome, their sharing behaviour negatively correlated with ToM abilities. This suggests that these children were able to understand the other's perspective, but purposely choose not to be altruistic in a situation that did not involve reciprocity (Cowell, Samek, List, & Decety, 2015). Conversely, when playing the ultimatum game, a situation that involves reciprocity, children with higher ToM abilities made fairer offers, suggesting that they were able to understand that unfair offers were more likely to be rejected (Takagishi, Kameshima, Schug, Koizumi, & Yamagishi, 2010).

Neuroscientific evidence is also inconclusive with respect to the directionality of the relationship between brain activation associated with mentalizing abilities and altruistic behaviour. For example, Chang and colleagues found that TPJ was more active when people decided to reciprocate trust of the investor in a trust game (see Box 1; Chang, Smith, Dufwenberg, & Sanfey, 2011); on the other hand, van Baar and colleagues reported that another ToM key area, the posterior superior temporal sulcus (pSTS), was more active when people chose to not reciprocate trust (van Baar et al., 2019). Similarly, some studies have found TPJ to be involved with the decision to react to unfair behaviour (David, Hu, Krüger, & Weber, 2017), whereas others found the area to be associated with the decision to refrain from punishing said behaviour (Stallen et al., 2018). Moreover, Buckholtz et al. (2008) investigated punishing decisions and responsibility assessment in a legal context and found TPJ to be involved in assessing all levels of criminal responsibility. In conclusion, mentalizing is clearly associated with the evaluation of the other's perspective, but not necessarily in a way that predicts the directionality of behavioural outcomes.

We will see in the next section how expectations shape social norms, and how mentalizing allows us to adapt these norms to different situations.

From Mentalizing to Social Norms

The Stanford Encyclopaedia of Philosophy defines social norms as 'the informal rules that govern behaviour in groups and society, [… and] the unplanned result of individuals' interaction' (Bicchieri, Muldoon, & Sontuoso, 2018). Social norms, such as fairness, cooperation, or trust, can therefore be interpreted as rules based on acquired expectations of the outcomes of social exchange, which have been learnt through repeated interactions. For this reason, mentalizing is crucial to the acquisition of social norms, as it is via this process that we are able to predict others' mental states and associated behaviours and, as a consequence, apply the correct social rule.

As mentioned above, social norms are rules that have been acquired through repeated exposure to social interactions. How does this learning happen? As proposed by Lee and Seo (2016), reinforcement learning theory can explain how we learn to respond to social tasks (i.e. apply social norms) that require decision-making. Model-free algorithms, where the likely outcome of each option of a decision task is compared to the value of the pre-decision state, can work in simple situations, such as when we have to choose between an apple and an orange. However, these algorithms are not suitable for complex and ever-changing environments such as social ones, where the many variables involved are constantly changing through time; this is because model-free algorithms require many iterations of events to update the response to any small change in the context, and therefore complex situations would require too many repeated interactions for the learning to take place. For this reason, model-free algorithms might not correctly capture the way in which social norms are acquired. Model-based algorithms, on the other hand, compute the value of each option based on the decision maker's knowledge of the situation, such as the other's beliefs, thoughts, and emotional state. Computationally, these model-based algorithms are more complex, but, thanks to their flexibility, are also more suitable to successfully describe and predict how we make decisions in the social environment, and therefore better explain how social norms are acquired. Mentalizing makes model-based algorithms of social decision-making psychologically feasible and, importantly, this is also the mechanism that distinguishes social and non-social learning and decision-making. Neurophysiological evidence supports this distinction, in that neural areas specific to the evaluation of the other's outcomes (TPJ, pSTS) are involved in updating social prediction-error (Suzuki et al., 2012); in other words, mentalizing processes contribute to shaping expectations on others' social behaviours.

The results from Heijne and Sanfey (2015) clearly illustrate this distinction between social and non-social learning. The authors investigated the mechanisms of decision-making in a stay/leave social situation, in which participants had to choose whether or not to leave either a social or a non-social partner in order to succeed in a cooperative task; two studies were run, one in which participants had no information about their partner, and one in which they were given prior knowledge to shape their beliefs. The findings showed that, as expected, the (non)cooperative behaviour of the partner influenced the decision to stay or leave the relationship. Prior beliefs also had an effect, biasing the decisions, though sometimes in a maladaptive way such as situations when beliefs about behaviour did not match actual behaviour: for example, when a partner was presented as cooperative, but their actual behaviour was non-cooperative, choosing to stay in the relationship would be considered maladaptive. Importantly, these results also showed that prior beliefs had a relatively weaker effect on the social choice compared to the non-social one: people used both their prior knowledge about their partner and their partner's actual behaviour in order to make a decision about whether or not to stick with that partner; on the other hand, in the non-social context, participants were much more driven by their prior expectations. This supports the idea that social value-based decision-making cannot be fully explained using simple non-social reinforcement learning and reward-prediction theory; other processes must be also accounted for, such as mentalizing, which allows us to understand that people might change their preferences.

Although being repeatedly exposed to other people's behaviours and mental states is a useful way to learn and follow social norms, observing another person's behaviour and predicting these behaviours are two different processes, with mentalizing playing a major role in the latter. Haroush and Williams (2015) found that non-human primates (rhesus monkeys) playing the prisoner's dilemma (see Box 1), a game in which mutual cooperation is required in order to achieve the best outcome for both players, would reciprocate both cooperative and non-cooperative choices; but, somewhat surprisingly, they found that a group of neurons in the dorsal anterior cingulate cortex (dACC) of the decision maker would specifically encode and predict the other monkey's decision to cooperate before the decision was shown (players were required to decide simultaneously). Therefore, when the monkey saw their opponent's decision before choosing an action themselves, they successfully reciprocated; however, when the other's decision was unknown, monkeys chose to cooperate significantly more, suggesting that cooperation was the default norm. Importantly, these neurons exclusively encoded the predicted cooperative choice of the other, not of oneself; moreover, they were sensitive to social context, firing more when the monkeys were in the same room rather than in separate rooms. These interesting results suggest that (1) engaging in mentalizing, rather than simply observing behaviour, may lead to higher levels of cooperation, possibly because compliance with social norms is highly expected; (2) social context is needed in order to trigger the social element of the decision process (Sanfey, Civai, & Vavra, 2015). The latter is also supported by findings showing that the willingness to engage in fair and altruistic, but costly, behaviours diminishes when the all parties are guaranteed anonymity, hence eliminating reputational concerns (Kurzban, DeScioli, & O'Brien, 2007).

Expectations in Flexible Social Environments

As previously mentioned, model-based algorithms of decision-making are better suited to deal with the unique complexity and flexibility of the social environment. Indeed, when it comes to predicting social norm compliance, expected behavioural outcomes vary dramatically depending on many different variables. For example, in order to determine whether an outcome is fair or unfair, intentionality plays a very important role. Findings showed that when responders in an ultimatum game know that the proposer was required to make an unfair offer (i.e. a 'no-alternative' condition), rejections of unfairness decrease significantly (Sutter, 2007). Interestingly, the anterior insula (AI), a key area in detecting social norm violations (Chang & Sanfey, 2011; Corradi-Dell'Acqua, Civai, Rumiati, & Fink, 2012), was more active when participants rejected unfair offers in the no-alternative condition, and accepted unfair offers in the fair-alternative condition (Güroğlu, van den Bos, Rombouts, & Crone, 2010), suggesting that the act of rejecting an unfair offer when there is no alternative, and accepting an unfair offer when the fair alternative is available, are both perceived as violations of a social norm. This indicates that fairness norms are context-dependent, and that, in the no-alternative condition, accepting unfair offers represents the social norm; therefore, AI here signals a behavioural deviation from the norm when rejecting unfairness. Noticeably, TPJ and mPFC are also more active when rejecting unfairness in the no-alternative condition, stressing the importance of mentalizing in adapting the norm to the context.

Other variables also influence our perception of fair outcomes: wealth and need, for example, are considered when deciding how to share resources, and people tend to prefer unequal outcomes that favour poorer groups (Tricomi et al., 2010); merit and effort are also integrated in evaluating context-dependent fairness: for instance, these variables determine the amount that people are willing to sacrifice in a dictator game where the amount of money to share is determined by the work of players (Frohlich, Oppenheimer, & Kurki, 2004).

Expectations regarding the type of environment in which a decision is made also play a fundamental role. Sanfey (2009) and Chang and Sanfey (2011) show that manipulating expectations of responders in the ultimatum game change the likelihood of rejecting unfair offers. Here, before playing the game, participants were led to believe that proposers would be either fair or unfair; as predicted, those who expected fair offers were much more likely to reject unfairness compared to those that expected unfair offers. As an extension to this, Vavra and colleagues show that not only the average, but also the variance of the expected distribution can influence participants' choices: specifically, the mean offer amount determined the threshold for accepting offers, whereas the variance of the offers determined how strictly participants adhered to this threshold (Vavra, Chang, & Sanfey, 2018).

Overall, these results stress some important aspects. What is considered as 'fair' changes depending on different contextual variables (e.g. intentionality, merit, effort), and our value-based decisions change accordingly (e.g. we prefer an unequal outcome if we know that our opponent needs the resources more than we do); however, even when our perception of a fair outcome remains the same (e.g. in Sanfey (2009) people would still consider the equal outcome, 50:50, to be a fair share), our prior expectations regarding the chances of obtaining the preferred outcome also can shift our decision threshold: this means that we may decide to accept an unfair offer, even though we would still prefer a fair one, if unfairness is what we expect in a specific context.

Individual Differences in Mentalizing and Social Norm Compliance

It is difficult to clearly understand the involvement of mentalizing in social value-based decision-making without considering individual differences. Quantifying the average behaviour of the population in specific social interactions can be very useful, for example to devise large-scale interventions such as social policies. Nevertheless, in order to understand the psychological roots underlying the multi-faceted and multidimensional decision-making mechanisms, it is important to investigate the complex interactions between individual and contextual variables. For instance, in an ultimatum game people on average prefer fairness and offer an equal split most of the time; however, as previously mentioned, the amount of money that each individual proposer chooses to offer will depend on their own beliefs about their game partner and the situation more generally. Typically, how people react to social norm violations involves the interaction of many different variables. Some of these variables are context-dependent, such as relative inequality of outcomes, reputation effects, need, merit, and anonymity. Other variables are more directly tied to the decision maker, such as age (Bailey, Ruffman, & Rendell, 2012; Murnighan & Saxon, 1998), gender (Solnick, 2001), or political beliefs (Zettler & Hilbig, 2010). As far as mentalizing and perspective taking are concerned, studies show that when required to choose between punishing a perpetrator or assisting a victim of an injustice, people with higher empathic traits show an increased disposition towards helping behaviour (Leliveld et al., 2012), with this attitude correlating with activation in TPJ (Hu, Strang, & Weber, 2015). In a recent study, Civai and colleagues show that people classified as compensators, based on their preference to compensate the victim of an injustice rather than punish the perpetrator in a punishment/compensation task (see Box 1), have a stronger activation in TPJ as compared to people classified as punishers, i.e. people who prefer to punish a perpetrator rather than compensate a victim (Civai et al., 2019). Somewhat in contrast to this, van den Bos and colleagues found that TPJ activation in a trust game was modulated by participants' subjective-value orientation: prosocial participants, who care about their own as well as the other's gain, showed a higher TPJ activation when defecting, whereas proself individuals, who focus only on their own gain ignoring the other's, showed this association when reciprocating, suggesting that more prosocial individuals attended more to the need of the others when defecting their trust (van den Bos, van Dijk, Westenberg, Rombouts, & Crone, 2009).

As previously mentioned, neither psychological nor neuroscientific evidence point to a clear relationship between mentalizing abilities and social preferences. However, moving from a localization view towards considering functional connectivity between regions can be a productive approach to explore this relationship (Vavra, van Baar, & Sanfey, 2017). For example, van Baar et al. (2019), in their version of the trust game, where participants must decide whether or not to reciprocate the trust of the investor, identified two types of reciprocators: those who reciprocate because they behave according to their own internal fairness norm (inequity-averse players), and those who take into account the investor's perspective, reciprocating in order to match the investor's expectations (guilt-averse players). The guilt-averse players show a stronger functional connectivity between TPJ and the vmPFC (as in Strombach et al., 2015) as compared to the other group, suggesting that players with this specific approach to social interactions (i.e. avoiding guilt) integrate the other's perspective into the final value calculation, whereas players that use other strategies do not.

Conclusions

Mentalizing is an essential mechanism which allows us to evaluate available choice options in a social context and then make a decision: this process allows individuals to integrate the perspective of others in an attempt to better predict each of the potential outcomes and, eventually, to select the optimal solution. Neuroscientific evidence supports the idea that a specialized neural circuit encodes this information, and that the derived signal is then integrated with other aspects into an overall value signal that informs the decision maker about the preferred option. Importantly, while the ability to mentalize and take the other's perspective is crucial in order to build a predictive model of the other's behaviour, it is not straightforward to correlate this to specific behavioural outcomes: for example, mentalizing and generosity are not always positively related. In order to understand the psychological roots of these mechanisms, individual differences must be taken into account to explain the multifaceted motivational drives that lead to the broad spectrum of behavioural outcomes.

To conclude, the ability to predict others' beliefs, emotions, and states of mind is fundamental to successful social decision-making; these observations have important implications when considering the effects that abnormal functioning of these mechanisms, either via brain damage or certain personality spectra, may have on people's ability to make optimal, or at least predictable, decisions. In the clinical setting, for example, suboptimal behaviour may get in the way of rehabilitation, preventing a good recovery unless different strategies are adopted (Bechara, 2005); in the forensic setting, the issue of criminal responsibility is tightly linked to the concept of mental ability, and therefore establishing the extent of this trait, and any contributing factors, is extremely important (Gazzaniga, 2008). Therefore, it is important to consider these implications in settings where abnormal behaviour needs to be explained and taken into account for successfully addressing the issues at hand.

Bibliography

Bailey, P. E., Ruffman, T., & Rendell, P. G. (2012). Age-related differences in social economic decision making: The ultimatum game. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 68*(3), 356–363.

Bechara, A. (2005). Decision making, impulse control and loss of willpower to resist drugs: A neurocognitive perspective. *Nature Neuroscience, 8*(11), 1458.

Benenson, J. F., Pascoe, J., & Radmore, N. (2007). Children's altruistic behavior in the dictator game. *Evolution and Human Behavior, 28*(3), 168–175. https://doi.org/10.1016/j.evolhumbehav.2006.10.003

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*(1), 122–142.

Bicchieri, C., Muldoon, R., & Sontuoso, A. (2018). "Social Norms", The Stanford Encyclopedia of Philosophy (Winter 2018 Edition), Edward N. Zalta(ed.), URL = <https://plato.stanford.edu/archives/win2018/entries/social-norms/>.

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron, 60*(5), 930–940. https://doi.org/10.1016/j.neuron.2008.10.016

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Chang, L. J., & Sanfey, A. G. (2011). Great expectations: Neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience, 8*(3), 277–284.

Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron, 70*(3), 560–572. https://doi.org/10.1016/j.neuron.2011.02.056

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics, 117*(3), 817–869. https://doi.org/10.1162/003355302760193904

Civai, C., & Hawes, D. R. (2016). Game theory in neuroeconomics. In *Neuroeconomics* (pp. 13–37). Berlin: Springer.

Civai, C., Huijsmans, I., & Sanfey, A. G. (2019). Neurocognitive mechanisms of reactions to second-and third-party justice violations. *Scientific Reports, 9*(1), 9271.

Clithero, J. A., & Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. Social cognitive and affective neuroscience, 9(9), 1289-1302.

Corradi-Dell'Acqua, C., Civai, C., Rumiati, R. I., & Fink, G. R. (2012). Disentangling self-and fairness-related neural mechanisms involved in the ultimatum game: An fMRI study. *Social Cognitive and Affective Neuroscience, 8*(4), 424–431.

Cowell, J. M., Samek, A., List, J., & Decety, J. (2015). The curious relation between theory of mind and sharing in preschool age children. *PLoS One, 10*(2), e0117947. https://doi.org/10.1371/journal.pone.0117947

David, B., Hu, Y., Krüger, F., & Weber, B. (2017). Other-regarding attention focus modulates third-party altruistic choice: An fMRI study. *Scientific Reports, 7*, 43024. https://doi.org/10.1038/srep43024

de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science (New York, N.Y.), 305*(5688), 1254–1258. https://doi.org/10.1126/science.1100735

Dixon, M. L., Thiruchselvam, R., Todd, R., & Christoff, K. (2017). Emotion and the prefrontal cortex: An integrative review. *Psychological Bulletin, 143*(10), 1033.

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior, 47*(2), 268–298. https://doi.org/10.1016/j.geb.2003.06.003

Edele, A., Dziobek, I., & Keller, M. (2013). Explaining altruistic sharing in the dictator game: The role of affective empathy, cognitive empathy, and justice sensitivity. *Learning and Individual Differences, 24*, 96–102.

Falk, A., Fehr, E., & Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry, 41*(1), 20–26.

Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: The neural circuitry of social preferences. *Trends in Cognitive Sciences, 11*(10), 419–427.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*(2), 63–87.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics, 114*(3), 817–868.

Frohlich, N., Oppenheimer, J., & Kurki, A. (2004). Modeling other-regarding preferences and an experimental test. *Public Choice, 119*(1), 91–117. https://doi.org/10.1023/B:PUCH.0000024169.08329.eb

Gazzaniga, M. S. (2008). The law and neuroscience. *Neuron, 60*(3), 412–415.

Güroğlu, B., van den Bos, W., Rombouts, S. A. R. B., & Crone, E. A. (2010). Unfair? It depends: Neural correlates of fairness in social context. *Social Cognitive and Affective Neuroscience, 5*(4), 414–423. https://doi.org/10.1093/scan/nsq013

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization, 3*(4), 367–388. https://doi.org/10.1016/0167-2681(82)90011-7

Haroush, K., & Williams, Z. M. (2015). Neuronal prediction of opponent's behavior during cooperative social interchange in primates. Cell, 160(6), 1233-1245.

Heijne, A., & Sanfey, A. G. (2015). How social and nonsocial context affects stay/leave decision-making: The influence of actual and expected rewards. PloS one, 10(8), e0135226.Hu, Y., Strang, S., & Weber, B. (2015). Helping or punishing strangers: Neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Frontiers in Behavioral Neuroscience, 9*, 24. https://doi.org/10.3389/fnbeh.2015.00024

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron, 87*(2), 451–462. https://doi.org/10.1016/j.neuron.2015.06.031

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron, 58*(2), 284–294. https://doi.org/10.1016/j.neuron.2008.03.020

Kang, P., Lee, J., Sul, S., & Kim, H. (2013). Dorsomedial prefrontal cortex activity predicts the accuracy in estimating others' preferences. *Frontiers in Human Neuroscience, 7*, 686.

Knutson, B., Westdorp, A., Kaiser, E., & Hommer, D. (2000). FMRI visualization of brain activity during a monetary incentive delay task. *NeuroImage, 12*(1), 20–27. https://doi.org/10.1006/nimg.2000.0593

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior, 28*(2), 75–84. https://doi.org/10.1016/j.evolhumbehav.2006.06.001

Lee, D., & Seo, H. (2016). Neural basis of strategic decision making. *Trends in Neurosciences, 39*(1), 40–48. https://doi.org/10.1016/j.tins.2015.11.002

Leliveld, M. C., van Dijk, E., & van Beest, I. (2012). Punishing and compensating others at your own expense: The role of empathic concern on reactions to distributive injustice. *European Journal of Social Psychology, 42*(2), 135–140. https://doi.org/10.1002/ejsp.872

McClure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2007). Time discounting for primary rewards. *Journal of Neuroscience, 27*(21), 5796–5804. https://doi.org/10.1523/JNEUROSCI.4246-06.2007

Morishima, Y., Schunk, D., Bruhin, A., Ruff, C. C., & Fehr, E. (2012). Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism. *Neuron, 75*(1), 73–79. https://doi.org/10.1016/j.neuron.2012.05.021

Murnighan, J. K., & Saxon, M. S. (1998). Ultimatum bargaining by children and adults. *Journal of Economic Psychology, 19*(4), 415–445.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience, 9*(7), 545.

Rangel, A., & Clithero, J. A. (2014). Chapter 8—The computation of stimulus values in simple choice. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics* (2nd ed., pp. 125–148). San Diego, CA: Academic Press. https://doi.org/10.1016/B978-0-12-416008-8.00008-5

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron, 35*(2), 395–405. https://doi.org/10.1016/S0896-6273(02)00755-9

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. Neuroimage, 22(4), 1694-1703.

Roy, M., Shohamy, D., & Wager, T. D. (2012). Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in Cognitive Sciences, 16*(3), 147–156.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience, 15*(8), 549.

Sanfey, A. G. (2009). Expectations and social decision-making: Biasing effects of prior knowledge on ultimatum responses. *Mind & Society, 8*(1), 93–107.

Sanfey, A. G., Civai, C., & Vavra, P. (2015). Predicting the other in cooperative interactions. *Trends in Cognitive Sciences, 19*(7), 364–365.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews, 42*, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009

Smith, D. V., Clithero, J. A., Boltuck, S. E., & Huettel, S. A. (2014). Functional connectivity with ventromedial prefrontal cortex reflects subjective value for social rewards. *Social Cognitive and Affective Neuroscience, 9*(12), 2017–2025. https://doi.org/10.1093/scan/nsu005

Solnick, S. J. (2001). Gender differences in the ultimatum game. *Economic Inquiry, 39*(2), 189–200.

Stallen, M., Rossi, F., Heijne, A., Smidts, A., Dreu, C. K. W. D., & Sanfey, A. G. (2018). Neurobiological mechanisms of responding to injustice. *Journal of Neuroscience, 38*(12), 2944–2954. https://doi.org/10.1523/JNEUROSCI.1242-17.2018

Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., & Kirsch, P. (2011). Beyond revenge: Neural and genetic bases of altruistic punishment. *NeuroImage, 54*(1), 671–680. https://doi.org/10.1016/j.neuroimage.2010.07.051

Strombach, T., Weber, B., Hangebrauk, Z., Kenning, P., Karipidis, I. I., Tobler, P. N., & Kalenscher, T. (2015). Social discounting involves modulation of neural value signals by temporoparietal junction. *Proceedings of the National Academy of Sciences, 112*(5), 1619–1624. https://doi.org/10.1073/pnas.1414715112

Sul, S., Tobler, P. N., Hein, G., Leiberg, S., Jung, D., Fehr, E., & Kim, H. (2015). Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. *Proceedings of the National Academy of Sciences, 112*, 201423895. https://doi.org/10.1073/pnas.1423895112

Sutter, M. (2007). Outcomes versus intentions: On the nature of fair behavior and its development with age. *Journal of Economic Psychology, 28*(1), 69–78.

Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., … Nakahara, H. (2012). Learning to simulate others' decisions. *Neuron, 74*(6), 1125–1137. https://doi.org/10.1016/j.neuron.2012.04.030

Takagishi, H., Kameshima, S., Schug, J., Koizumi, M., & Yamagishi, T. (2010). Theory of mind enhances preference for fairness. *Journal of Experimental Child Psychology, 105*(1–2), 130–137.

Tricomi, E., Rangel, A., Camerer, C. F., & O'Doherty, J. P. (2010). Neural evidence for inequality-averse social preferences. *Nature, 463*(7284), 1089–1091. https://doi.org/10.1038/nature08785

van Baar, J. M., Chang, L. J., & Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature Communications, 10*(1), 1483.

van den Bos, W., van Dijk, E., Westenberg, M., Rombouts, S. A., & Crone, E. A. (2009). What motivates repayment? Neural correlates of reciprocity in the Trust Game. *Social Cognitive and Affective Neuroscience, 4*(3), 294–304.

Vavra, P., Chang, L. J., & Sanfey, A. G. (2018). Expectations in the ultimatum game: Distinct effects of mean and variance of expected offers. *Frontiers in Psychology, 9*, 9. https://doi.org/10.3389/fpsyg.2018.00992

Vavra, P., van Baar, J., & Sanfey, A. (2017). The neural basis of fairness. In M. Li & D. P. Tracer (Eds.), *Interdisciplinary perspectives on fairness, equity, and justice* (pp. 9–31). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-58993-0_2

Zettler, I., & Hilbig, B. E. (2010). Attitudes of the selfless: Explaining political orientation with altruism. *Personality and Individual Differences, 48*(3), 338–342. https://doi.org/10.1016/j.paid.2009.11.002

Mentalizing in Value-Based Vicarious Learning

Lisa Espinosa, Armita Golkar, and Andreas Olsson

Introduction

Humans routinely learn the value of things and actions by observing and interacting with others. Such vicarious learning experiences play a fundamental role in shaping our behavior across a range of situations, from simple avoidance responses when in danger to culturally specific actions in social contexts. What role does attributions of mental states—"mentalizing"—about the thoughts and feelings of others, play in such vicarious learning? And, how is such social learning computed by the brain? These are central questions discussed in this chapter. Surveying research across phylia, we conclude that mentalizing is important in several specific ways in human vicarious learning, but not a necessary prerequisite for vicarious learning to occur as illustrated by, for example, its ubiquity in the animal kingdom. Research in humans, which is our focus here, shows that vicarious learning is realized through the joint action of networks of brain regions responsible for mentalizing and domain-general learning processes.

Across many species, learning the value of stimuli and behaviors through observation is demonstrated early on in life. Toddlers, for example, quickly learn to express defensive responses and avoidance towards novel toy animals previously paired with negative facial expressions by their mothers (Gerull & Rapee, 2002), and children readily learn the value of arbitrary (Marshall & Meltzoff, 2014;

L. Espinosa · A. Olsson (✉)
Division of Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Solna, Sweden
e-mail: lisa.espinosa@ki.se; andreas.olsson@ki.se

A. Golkar
Division of Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Solna, Sweden

Department of Psychology, Stockholm University, Stockholm, Sweden
e-mail: armita.golkar@psychology.su.se; Armita.golkar@ki.se

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_26

Repacholi & Meltzoff, 2007), aggressive (Bandura, 1978), and avoidance (Askew & Field, 2007) behaviors by watching unknown adults, and later creatively express this learning in new contexts. Throughout their lifespan, humans continue to share information and socially learn about the value of things, people and their actions, enabling, for example, useful differentiations between friendly minded and hostile individuals, and culturally appropriate and inappropriate actions. The transmission of information from parents to children, as well as between peers, constitutes a core mechanism of adaptive cultural learning (Boyd & Richerson, 2009; Tomasello, 2011), but can also cause maladaptive behaviors, such as anti-social actions, exaggerated avoidance, and anxiety (Bandura, 1978; Debiec & Olsson, 2017; Hopwood & Schutte, 2017).

In our species, vicarious learning often involves some kind of inferences about the content of the minds of the people we are learning from ("demonstrators"). Thinking about others' minds or "mentalizing" includes updating inference about demonstrators' intentions, thoughts, and feelings (see Chap. 30). For example, watching an individual displaying defensive behaviors typically associated with fear and anxiety in a threatening context, say when attacked by a mob, might lead you to attribute the experience of fear to the individual, accompanied by an empathic feeling of distress. In turn, these processes may critically affect how and what you learn from your experiences. You might, for example, learn to avoid the location where the assault took place, and dislike and distrust people from the social category you are ascribing to the mob. Based on your impressions, you might also simulate behavioral strategies to escape, and these strategies might be useful if you happen to find yourself in a similar situation in the future. Should you instead interpret the situation as playful, and the expressions of the target individual as excitement, your learning might be radically different. In the vicarious situations described above, mentalizing about, and affectively sharing, the demonstrator's subjective states jointly contribute to the observer's empathic response. Importantly, mentalizing and affective resonance are thought to rely on different neural processes engaging network of regions computing reflective, cognitive and self-experienced, affective qualities, respectively (Zaki & Ochsner, 2012). It should be noted that just as the attributed content of others' minds can serve as motivating and informative states of the world, thus providing the basis for learning, our attributions and of mental states to others, can themselves also be modified by learning experiences, for example through error correcting mechanisms (Hein, Engelmann, Vollberg, & Tobler, 2016; Lockwood, Apps, Valton, Viding, & Roiser, 2016; Olsson & Spring, 2018), which will be discussed in greater extent in later sections.

Demonstrations of vicarious learning in young individuals without fully developed mentalizing abilities (Gerull & Rapee, 2002), and the expression of vicarious learning without conscious awareness (Olsson & Phelps, 2004), suggest however that mentalizing is not necessary for such learning to occur in our species. Similarly, other social influences closely related to vicarious learning, but not discussed in the present chapter, such as imitation, conformity-based decision-making, and local enhancement, might not require (although often influenced by) mentalizing processes.

To empirically examine different aspects of vicarious forms of learning, researchers have used a range of simple experimental models aiming to capture the essential elements of the learning situation. One such commonly used model exposes participants ("observers") to pre-recorded (Fig. 1; Haaker, Golkar, Selbing, & Olsson, 2017) or live (Pärnamets, Espinosa, & Olsson, 2019) demonstrators that undergo direct Pavlovian threat conditioning. During the learning stage, both the observer and demonstrator view initially neutral conditioned stimuli (CS), some of which are occasionally paired with a direct aversive event, such as an electric shock, to the demonstrator (but not the observer). The efficacy of the vicarious learning is assessed at a later time (during "test") by measuring the observer's learned responses to the CS in the absence of the demonstrator. As will be discussed in detail later, learning from such social experiences is dependent on mental attributions to the demonstrator.

A related line of research has extended the vicarious analogue to learning about safety (Golkar, Castro, & Olsson, 2015; Golkar, Haaker, Selbing, & Olsson, 2016; Golkar & Olsson, 2016; Golkar, Selbing, Flygare, Öhman, & Olsson, 2013). In these experiments, participants watch a calm looking demonstrator modeling safety when presented with a CS that the participant previously learned to associate with a direct aversive experiences, such as a mild shock to the wrist. This line of research has revealed that vicarious safety learning leads to superior attenuation of the conditioned threat response in comparison to traditional direct safety learning training when no demonstrator is present (Golkar et al., 2013). Interestingly, the efficacy of vicarious safety learning depends on the demonstrator being depicted as calm (Golkar et al., 2013, 2016), suggesting that the attribution of subjective safety to the demonstrator is critical for successful downregulation of threat responses by means of social observation.

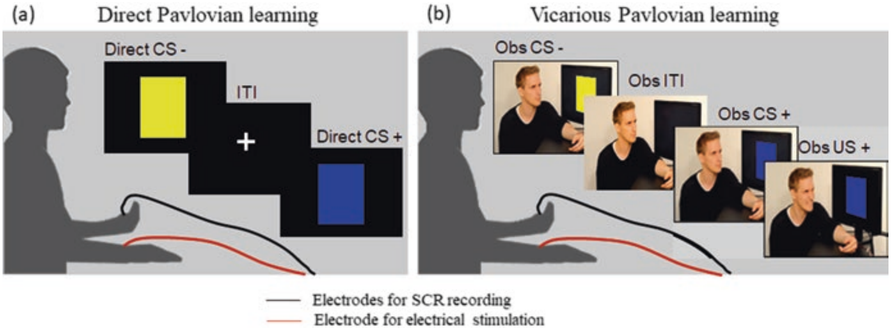

(a)
Direct Pavlovian learning
Direct CS -
ITI
+
Direct CS +
(b)
Vicarious Pavlovian learning
Obs CS -
Obs ITI
Obs CS +
Obs US +
Electrodes for SCR recording
Electrode for electrical stimulation

Fig. 1 Overview of vicarious Pavlovian learning protocol using skin conductance responses (SCR) to index learning. (**a**) The *Learning stage* depicts the observer (participant in shaded gray) watching the demonstrator's responses to receiving a shock paired one out of two colored squares serving as conditioned stimuli (CS); and (**b**) The *Test stage* illustrates the participant being directly exposed to the CS (note that no shocks are administered to the participant in any stage). (*Figure modified from* Haaker, Golkar, et al. (2017) and Haaker, Yi, et al. (2017))

In spite of the surge in research on various psychological and neural aspects of social learning over the past decades, surprisingly little is known about the involvement of mentalizing, and the role of building internal mental models of others' minds, during these processes. This is the focus of the current chapter. The lion's share will be dedicated to discussing mentalizing in vicarious learning transmitted through observation. Although the emphasis will be on the vicarious analogue of traditional stimulus-stimulus learning, here referred to as "vicarious Pavlovian" learning, we also discuss developments related to "vicarious instrumental" learning. In particular, we will survey recent research applying reinforcement learning models to better understand the computational aspects of learning to value own and others' actions through observational experiences. Although we will cover research on both appetitive and aversive social learning, the focus will be on aversive processes.

Connecting Vicarious Emotions, Mentalizing, and Learning: Historical Perspectives

Vicarious emotions and their impact on learned behavior have been described by philosophers and used in the arts since the origin of the Greek tragedies. In modern times, British Enlightenment philosophers David Hume (1711–1776) and Adam Smith (1723–1790) argued that vicarious emotions ("passions") are critical to the individual's social and moral development, and therefore to a well-functioning society. Although their understanding of the mechanisms was rudimentary, Hume's speculations about the processes underlying vicarious emotions were prophetic for contemporary research on vicarious learning, including both the ideas of emotional "mirroring" and "associations" between the present situation and earlier encounters of similar events (Hume, 1985). Critically, these early theorists also noted that the interpretation of others' emotional expressions (mentalizing) shapes their impact of social observations on the observer.

The systematic investigation of the role of mentalizing in vicarious learning goes back to early experimental work in humans on "vicarious emotional instigation" referring to the inference of a demonstrator's unconditioned response (UR) following the presentation of an aversive unconditioned stimulus (US) (Berger, 1962; Hygge & Öhman, 1978; Lanzetta & Englis, 1989). For example, a seminal study by Berger (1962) showed that another person's arm movement in response to an alleged shock instigated threat learning in the observer, but only when the observer believed that the movement was caused by a shock, and not when the demonstrator's arm moved without a shock or when a shock was delivered without movements of the arm. Other early work showed that the appraisal of a demonstrator's internal emotional state after noticing changes in the demonstrator's heart rate (Kravetz, 1974), as well as information about another person's spider phobia, but without observing any overt responses (Hygge & Öhman, 1978), can induce learned threat responses to the phobic stimulus (a picture of a spider). These, and other similar findings, together with clinical observations of the social transmission of disruptive fears and anxieties, led to an influential theoretical model of how fears and phobias can develop through vicarious transmission (Askew & Field, 2007; Rachman, 1977).

The findings reviewed so far show that attributions of mental states to the demonstrator determine the quality of the ensuing vicarious learning in our species. Social learning is, however, common in many animals. For example, a long line of research on observational learning across animals, from rodents (Chang & Debiec, 2016; Jones, Riha, Gore, & Monfils, 2014; Knapska, Mikosz, Werka, & Maren, 2010; Monfils & Agee, 2019) to non-human primates (Chang, Gariépy, & Platt, 2013; Cook & Mineka, 1989), has verified the efficiency of this social route to learning in many different ecologies, and described its neural underpinnings in great detail. Importantly, these findings show that vicarious threat learning does not require human level of sophisticated social cognition. It remains an open question, however, what minimal cognitive capacities are necessary for this form of learning to occur: internal simulations (Goldman, 1992; Hesslow, 2002; Rizzolatti & Craighero, 2004), an internal model or theory (Theory of Mind), or simply domain-general associative learning mechanisms (Heyes, 2012; Olsson, Knapska, & Lindström, in press). In this chapter, we discuss research that addresses this question by surveying various forms of social learning in humans with both intact and impaired social cognitive abilities, as well as a selection of non-human species that may inform our understanding of the role of various forms of mentalizing in vicarious learning. Next, we survey domain-general learning processes that seem to be shared between learning through direct and vicarious experiences, and how these processes interact with social cognition.

From Direct to Vicarious Pavlovian Learning

A common model of emotional learning is that of Pavlovian associative conditioning (Pavlov, 1927; Phelps & LeDoux, 2005), in which an individual's direct experiences of predictive pairings of a neutral conditioned stimulus (CS) and a naturally aversive or rewarding unconditioned stimulus (US) result in the expression of a conditioned response (CR). Research across species has strengthened the suggestion that these associative mechanisms provide a foundation for various forms of social learning, such as vicarious learning about threats (Cook & Mineka, 1989; Debiec & Olsson, 2017) and rewards (Morelli, Knutson, & Zaki, 2018; Seymour, Singer, & Dolan, 2007). Support for this assertion comes from findings using behavioral and neural measures, as well as computational models of learning. For example, a seminal study in monkeys (Cook & Mineka, 1989) showed that the relationships between the strength of expressed distress in a demonstrator, the observer's immediate response to the demonstrator's behavior, and the resulting threat learning in the observer, corresponded to the well-established relationship reported between US, UR, and CR in direct Pavlovian conditioning. The same relationship has been described in humans (Debiec & Olsson, 2017; Olsson, Nearing, & Phelps, 2007), but not in rodents (Kavaliers, Choleris, & Colwell, 2001), suggesting a greater reliance on expressivity during the learning process in primates. Indeed, the musculature of the primate face allows it to produce a greater variety of emotional expressions, superior to that of many other species (Ekman, 1982), and the cortical areas dedicated to face processing are also enlarged in primates relative to other species (Rolls, 1999). Among primates, the richness and flexibility of the human face is unparalleled, allowing for a greater range of expressions. In addition, the structural connectivity between face processing regions in the visual cortex and the amygdala, known to be critical for direct threat learning across species, is vastly developed in humans (Bickart, Dickerson, & Feldman Barrett, 2014; Rolls, 1999), supporting the unique sophistication of both the expression and decoding of facially transmitted emotional information in our species. The function of the demonstrator's expressions in vicarious learning has led to it being referred to as a "social US" (Debiec & Olsson, 2018; Olsson & Phelps, 2007; Olsson et al., in press), which might have an intrinsic (non-learned) or learned emotional meaning, or both. This implies that social cues, like other Pavlovian cues, can exert their influence without higher cognitive processes, such as expectations and mentalizing. Indeed, this is supported by demonstrating that social and non-social Pavlovian cues have partially overlapping, behavioral (Askew & Field, XXX), computational (Lindström, Golkar, Jangard, Tobler, & Olsson, 2019), and neural (Olsson, Nearing & Phelps; Olsson et al., in press) characteristics. Although social US, such as a fearful face, can play a similar role as, for example, the direct experience of a shock (US), the dynamic and variably expressive human face and other social cues tied to the demonstrator provide focal targets for mentalizing, which is the topic discussed next.

Mentalizing in Vicarious Pavlovian Learning

The historical focus on non-social learning based on direct experiences in learning research, and on mentalizing processes in isolation from their role in learning in social psychology, has resulted in a dearth of knowledge about the mechanisms of vicarious learning in general, and of mentalizing in this kind of learning, in particular. As described above, it is reasonable to assume that most human vicarious learning involves social cognition, such as perceptions and interpretations of physical expressivity, prior social knowledge and expectations. The updated attributions about thoughts and intentions of others play an important role in understanding and predicting the behavior of others, as well as learn from it (Frith & Frith, 2012; Olsson & Ochsner, 2008), thus providing an important selective pressure for the evolution of mentalizing abilities (Tomasello, 2011). Attributions can be based on observable cues (e.g., responses and actions), as well as hidden factors (transient mental states and stable traits), all of which are used to understand and predict others' behaviors (Tamir & Thornton, 2018). Importantly, dynamically evolving mental state and trait attributions are affected by our prior knowledge and valuations of individuals and specific social groups. Next, we discuss how social cues, for example, markers of social group and dominance, affect vicarious learning..

Social regulation of vicarious Pavlovian learning. Social cues in the environment can modulate how we perceive and judge others, providing indirect evidence for the involvement of mentalizing about the demonstrator's thoughts and dispositions during vicarious learning. A number of behavioral studies have demonstrated the influence of such cues, as well as the specific relationship between the observer and demonstrator, on social threat and safety learning. A first example of this is the role played by identity. For example, vicarious learning has shown to be improved when threat and safety information is transmitted from individuals belonging to the same racial (Golkar et al., 2015) and culturally determined (Golkar & Olsson, 2017) group. Pointing towards a possible motivational basis, subjective ratings of social group identification was positively related to vicarious learning from in-group demonstrators (Golkar & Olsson, 2017) and negative racial attitudes were negatively related to learning from an out-group demonstrator (Golkar et al., 2015). Similarly to research in humans, improved vicarious learning has demonstrated between familiar and genetically related rodents (Jones et al., 2014; Kavaliers et al., 2001) and non-human primates (Apps, Rushworth, & Chang, 2016; Chang et al., 2013). This does of course imply neither that higher-order order processes are involved in other animals, nor that mentalizing is not involved in humans. Indeed, similar behavioral effects can result from different underlying processes. For example, the bias demonstrated in humans might result from both altered basic learning properties, including stimulus saliency, and the integration of higher-order biased mentalizing. A second example is research on the effects of attributed dominance on vicarious learning. Following a task where participants were asked to attribute dominance to (the facial picture of) one individual over another after observing their dyadic confrontation, participants underwent a Pavlovian conditioning using these pictures as CS+. The results showed that although both facial pictures were equally predictive of a mild electric shock, the results indicated a stronger conditioned threat response to the picture of the dominant individual compared to the picture of the submissive individual (Haaker, Molapour, & Olsson, 2016). Because these studies did not directly examine mentalizing, they cannot make any firm conclusions to what extent such processes contributed to the results beyond domain-general associative learning processes. Yet, self-report measures suggest that attributions related to the perception of group identity, as well as dominance, contributed to vicarious learning.

The growing body of research described above indicates that social characteristics of the demonstrator or the situation can moderate how we learn about threat and safety. A possibility is that this effect is dependent on processes related to attention and/or affective sharing. For example, attention to perceived (or imagined) emotional expressions of the demonstrator are likely to facilitate emotional appraisals and physiological resonance with the demonstrator. In support of this, recent research has demonstrated that enhanced neural alignment of physiological responses in two individuals is related to better skill learning from (Pan, Novembre, Song, Li, & Hu, 2018) and liking of (Parkinson, Kleinbaum, & Wheatley, 2018) each-other. These findings of alignment are consistent with the conjecture that mentalizing, as well as affective sharing, contribute to vicarious learning. Addressing the specific role of affect sharing in vicarious threat learning directly, a recent study measured spontaneous synchronization of phasic skin conductance between observer and demonstrator during a live dyadic vicarious learning paradigm. The results demonstrated that the degree of synchronicity, as well as self-reported empathy with the demonstrator, predicted the strength of the observer's conditioned response at later test (Pärnamets et al., 2019). The coupling of autonomic nervous systems between observer and demonstrator described here is informative about the processes underlying vicarious learning. Yet, to better understand the role of mentalizing, we need to consider research using instructed appraisals and perspective-taking, which is discussed next..

Appraisals and perspective-taking during vicarious Pavlovian learning. Appraising the content of others' minds and taking their perspectives are core psychological processes of mentalizing, and have shown to enhance vicarious emotional responding in an observer (Lamm, Batson, & Decety, 2007; Shu, Hassell, Weber, Ochsner, & Mobbs, 2017). In turn, these processes are likely to impact how we learn from a demonstrator. Importantly, motivational aspects play important roles in determining our responses to others' misery. For example, the pain of a demonstrator believed to be a future competitor has been shown to trigger the opposite responses in the observer: "schadenfreude" (pleasure in another's pain; Lanzetta & Englis, 1989). Similar effects accompanied with an attenuation of activity in brain regions linked to the affective components of empathy, the AI and the ACC, were shown when the target person was known by the observer to have been cheating in a previous economic game (Singer et al., 2006). These and more recent findings (Hein et al., 2016) have shown that the motivational significance of others' misery matter more for their role in learning than the measurable behavioral characteristics of those suffering. It should be noted that although an observer expresses no empathy, and maybe even schadenfreude, he or she might accurately attribute a negative internal state to the demonstrator. In other words, depending on the circumstances, mentalizing can be dissociated from affective responses.

To our knowledge, to date, there is only one study that has directly manipulated mentalizing to examine its role in vicarious Pavlovian learning (Olsson et al., 2016). This study upregulated and downregulated participants' emotional responses to a demonstrator in pain through a standard perspective-taking technique (Batson, Early, & Salvarani, 1997). The results revealed that encouraging state empathy by means of perspective-taking improved vicarious threat learning as measured during a later test in the absence of the demonstrator. Although this study demonstrated the direct importance of mentalizing in vicarious Pavlovian learning, evidence for the role of trait empathy in vicarious learning is mixed. Some studies show weak or no correlations (Olsson et al., 2016; Williams & Conway, 2019), and others report a positive relationship (Kleberg, Selbing, Lundqvist, Hofvander, & Olsson, 2015;

Lockwood, Apps, Roiser, & Viding, 2015). These discrepancies might partially be explained by the heterogeneity of scales used to capture individual differences of a concept that itself is debated and often ill defined. Interestingly, a recent study examining vicarious threat learning in individuals diagnosed with autism spectrum disorder, a group typically associated with low trait empathy, reported enhanced learning as compared to a matched healthy control (Espinosa et al., in press). This finding might be explained by an enhanced attention towards, and an impaired regulation of, the vicarious threat in this neuroatypical group.

In sum, research on vicarious Pavlovian learning has shown that a range of manipulations of the social characteristics of the demonstrator affect the learning outcome. Most of these studies have not directly manipulated mentalizing in the observer during learning, but the results convincingly support the role of inferences about others' thoughts, feelings, and dispositions, as well as affective sharing processes. Described with the terminology introduced earlier, the demonstrator's emotional expressions seem to function as a social US, and the strength and meaning of this can be regulated by a range of social cognitive factors. Additional support for the same conclusion is provided by research on the neural substrates of these processes, which will be discussed next.

Extending the Neural Model of Direct Learning to Social Learning

Most of our knowledge about the neurobiological mechanisms of threat learning comes from the study of direct Pavlovian learning (LeDoux, 2000). This neural model has subsequently been extended to explain social threat learning, in particular vicarious Pavlovian learning through observation. Accordingly, brain regions implicated in domain-general learning and affective processes interact with those involved in social cognition (Debiec & Olsson, 2017; Olsson & Phelps, 2007): First, the "core aversive learning network," partially independent from higher cognitive functions, is centered on the amygdala, a region critical for the acquisition, storage, and expression of direct conditioning. This network also includes the ACC and AI that compute (self and others') evaluations and subjective experiences (Haaker, Yi, Petrovic, & Olsson, 2017; Lindström, Haaker, & Olsson, 2018; Meffert, Brislin, White, & Blair, 2015; Olsson et al., 2007). The "social cognitive network" supports the processing of information regarding (self and others') attributions of mental states, and mobilizes a network of regions, including among others, the superior temporal sulcus, STS (Carlin & Calder, 2013), tempo-parietal junction, TPJ (Saxe & Wexler, 2005; Zaki & Ochsner, 2012), and the dorsal medial prefrontal cortex (MPFC; Ochsner et al., 2004; Zaki & Ochsner, 2012).

Computational approaches also support that vicariously and directly experienced learning rely on partly overlapping mechanisms. More precisely, research has described the contribution of domain-general principles, such as that of prediction error that provides an update of the knowledge about the world based on the difference between expected and actual events, regardless if these are vicarious or directly experienced (Joiner, Piva, Turrin, & Chang, 2017; Lindström et al., 2019; Ruff & Fehr, 2014). Computational approaches have been particularly common in research on vicarious forms of instrumental learning, which has implicated a set of brain regions partially different from Pavlovian learning, including dopaminergic projections from the ventral tegmental area (VTA) to the ventral striatum and prefrontal cortex (PFC) (Glimcher, 2011). Vicarious instrumental learning and its underlying computations will be discussed in more detail in the last section of this chapter.

Taken together, the research discussed so far has demonstrated that vicarious and direct Pavlovian threat learning are subserved by several common neurocomputational mechanisms. Importantly though, vicarious learning is distinguished by its processing of social information. Next, we will extend the discussion about the neural basis of mentalizing in vicarious learning, beginning with research in animals that includes the processing of social information lacking the sophistication of human mentalizing.

Neural Basis of Mentalizing in Vicarious Learning

Social cognition in non-human vicarious Pavlovian learning. Animal-based models have enhanced our understanding of the neural correlates of primal forms of empathy, mentalizing, and vicarious experience by allowing cross-species comparisons. Rodent research showed similar commonalities with human studies regarding the involvement of the amygdala both in direct and vicarious Pavlovian learning (Debiec & Sullivan, 2014; Jeon et al., 2010). These animal-based models demonstrated the involvement of the amygdala, ACC, PFC, and thalamic and hypothalamic nuclei in socially transferred emotions (Meyza, Bartal, Monfils, Panksepp, & Knapska, 2017). Pharmacological inactivation of the ACC and ACC-amygdala projection in mice was shown to affect acquisition during vicarious learning, while the same inactivation did not affect threat responses during direct learning (Jeon et al., 2010). These results were corroborated and extended by more recent work identifying specific pathways relaying social information from the ACC to the basolateral amygdala during vicarious learning in mice (Allsop et al., 2018). These findings suggest that vicarious learning relies on a hierarchical pathway in which socially derived aversive cue information is transmitted to lower-level regions involved in domain-general associative learning and defensive responding. The involvement of similar network of regions in vicarious learning in both non-human and human animals implies shared basic abilities at the core of social learning, which could be the foundation of the more complex cognitive abilities found in humans. Interestingly, studies in infant rats have shown that mother-to-infant social transmission of threat to a novel odor is accomplished by the elevation of the infant's corticosterone levels induced by the mother's frightened reaction to the novel odor (Debiec & Sullivan, 2014). This learning was mediated by the amygdala, and demonstrated by inactivating the lateral and basal nuclei of the amygdala in the infant rats, which prevented the mother-to-infant transmission of threat. A follow-up study extended these findings and demonstrated that unlike direct Pavlovian learning, which fully emerges during the infant's second week of life (Sullivan, Landers, Yeaman, & Wilson, 2000), vicarious learning is present at birth, allowing them to acquire long lasting threat responses from their mother before complete maturation of neocortical structures such as ACC, insular cortex (IC), or prelimbic cortex (Debiec & Sullivan, 2014). This pattern of findings suggest unique neural mechanisms supporting social learning in the developing brain in non-human animals. These findings again show that the sophisticated mental abilities present in humans are not necessary for observational learning. Yet, they might also provide important clues towards the role of social cognition, including mentalizing, in social learning.

Mentalizing in human vicarious Pavlovian learning. Accumulating evidence shows that higher-order cognitive processes, such as perspective-taking and mental attributions, influence the observer's learned response. In other words, the observer's learning is mediated by the observer's perception of the demonstrator. The social emotional learning model (Olsson & Phelps, 2007) described above suggested that vicarious emotional learning is distinguished from direct Pavlovian learning by its involvement of prefrontal social-cognitive network, including distinct activations of the aversive learning network, as well as its interactions with regions processing social information, among them the STS, dmPFC, and the TPJ.

The STS is a multimodal region integrating information from the action perception stream, including others' gaze direction (Carlin & Calder, 2013) and mentalizing processes carried out in the dmPFC (Amodio & Frith, 2006; Denny, Kober, Wager, & Ochsner, 2012). The tracking of motion, such as others' facial movements, is suggested to support implicit mentalizing by tracking intentions in others (Frith & Frith, 2012). Along with STS, the TPJ directs attention to salient information, represents others' beliefs, and has been causally linked to strategic mentalizing processes (Hill et al., 2017). This network of regions have been hypothesized to be involved in mentalizing about self and others' mental states (Adolphs, 2008; Amodio & Frith, 2006; Olsson & Ochsner, 2008), strengthening the assumption that vicarious learning involves regions is linked to perspective-taking (Fig. 2).

A study by Lindström et al. (2018) used dynamic causal modeling (DCM) to describe the flow of information in the amygdala-AI-ACC network during direct and vicariously experienced US (i.e., shock to the self, and to the demonstrator). The results demonstrated that information about the US was most likely to enter the network through the amygdala during direct learning and through the AI during vicarious learning. Moreover, participants' self-reported empathy with the demonstrator and the unpleasantness of observing the demonstrator receiving a shock correlated with activity in the AI during the observation stage. The involvement of the AI in vicariously experienced pain dovetails with the role of the AI and ACC in affective sharing and empathy (Adolphs, 2008; Lamm, Decety, & Singer, 2011; Zaki & Ochsner, 2012) and contributes to the explanation of why these brain regions (Olsson et al., 2007) and empathic appraisals (Olsson et al., 2016) have been shown to predict the strength of vicarious threat learning.

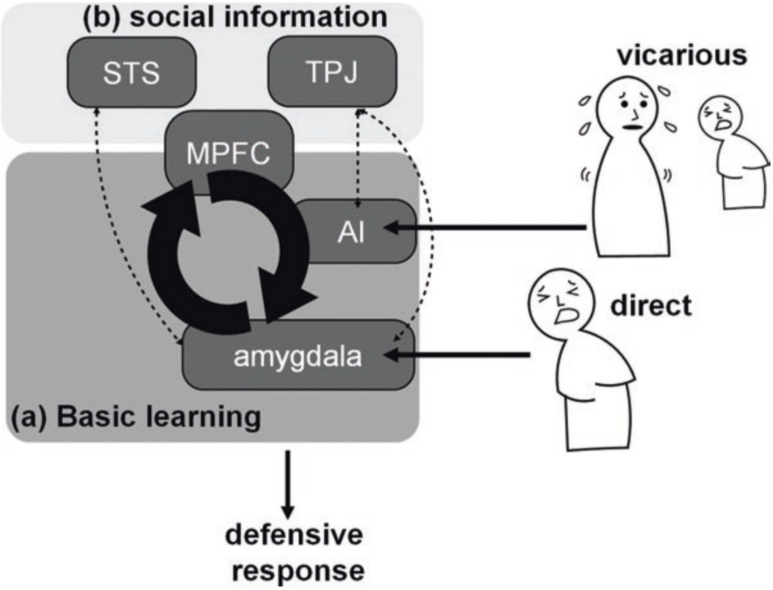


Fig. 2 Schematic illustration of a selection of interactive neural regions involved in threat learning through direct and vicarious means (adapted from Olsson et al., 2018). The dark grey area (**a**) indicates a network of regions involved in basic aversive learning, and the light grey area (**b**) includes regions involved in the processing of social information, e.g., mentalizing. The MPFC (here including the ACC) is highlighted in both networks because its subregions have been implicated in mentalizing (BA 9, Ochsner, 2004), learning from others' experiences (BA 32, Apps et al., 2016; Lockwood et al., 2015), as well as direct Pavlovian threat learning (BA 13 and 32, Fullana et al., 2016). Bold arrows to the right indicate likely inputs of vicarious and direct information, respectively (Lindström et al., 2018). Circular bold arrows refer to connectivity between three core regions during both direct and vicarious threat learning, and dashed lines indicate connectivity during threat learning (Lindström et al., 2018). *MPFC* medial prefrontal cortex, *ACC* anterior cingulate cortex, *AI* anterior insula, *STS* superior temporal sulcus, *TPJ* temporoparietal junction, *BA* Brodmann's area. (*Figure adapted from* Olsson, FeldmanHall, Haaker & Hensler, 2018)

In sum, research on Pavlovian vicarious learning supports the importance of both domain-general neural processes, subserving associative learning and attention, and social cognition. Human imaging research taken together with behavioral studies directly manipulating and measuring various forms of mentalizing, suggest that the demonstrator can serve as a social unconditioned stimulus (US). The fact that vicarious Pavlovian learning occurs across species, involving neural systems involved in relaying social information, suggests that there are many routes to successful observational threat learning. Indeed, social learning might be species-specific and highly dependent on the ecology in which it has evolved (Kendal et al., 2018).

So far, we have surveyed research targeting vicarious Pavlovian (stimulus-stimulus) learning. Next, we turn to the role of mentalizing in the reinforcement of behaviors through vicarious instrumental learning.

Vicarious reinforcement and instrumental learning in humans. Emotional expressions in the demonstrator do not only imbue stimuli and contexts with a value, i.e., Pavlovian vicarious learning. Analogous to personally experienced instrumental learning, the consequence of a demonstrator's behavior can also serve as a vicarious reinforcer of the same behavior in the observer. For example, rewarding a demonstrator's aggressive or prosocial actions reinforces the same behavior in the observer. Similar to vicarious Pavlovian learning, social cognition and mentalizing play important roles in the social learning of the value of actions. Notably, the vast majority of studies focusing on vicarious instrumental learning have targeted reward learning in simple decision-making tasks, as compared to vicarious Pavlovian learning studies, which have mainly targeted aversive learning to stimuli that are presented to a passive observer. Research on instrumental forms of vicarious learning has implied partially different brain regions as compared to vicarious Pavlovian learning. Similar to direct instrumental learning and decision-making, vicarious instrumental learning has been associated with dopaminergic projections from the VTA to the ventral striatum and the PFC (Glimcher, 2011). These dopaminergic projections have been associated with computational features of adjusting one's own behavior based on positive or negative reinforcement (also known as Reinforcement Learning, RL) to both self and others (Burke, Tobler, Baddeley, & Schultz, 2010; Joiner et al., 2017). A key feature of RL that has been linked to the dopaminergic system is prediction error, the computation of the difference between expected and actual outcomes of an action, such as a choice.

Importantly, reinforcement learning is a framework that has been increasingly used to describe learning and learned outcomes through mathematical models taking into account internal states, such as motivation and subjective value (Montague, Hyman, & Cohen, 2004). The RL framework can capture complex behaviors with relatively simple yet powerful models, such as the Rescorla–Wagner model (Rescorla & Wagner, 1972), to describe both direct and vicarious learning (Burke et al., 2010; Lindström et al., 2019). Although RL models come in different forms, they share key elements, such as a description of the rate of learning and salience of stimuli in order to fit the specifics of learning and decision-making processes.

Mentalizing during vicarious reinforcement. Applying computational models to mentalizing processes is a relatively recent research effort, enabling researchers to go beyond *where* in the brain social information is processed to making claims about *how* it is processed (Carter, Bowling, Reeck, & Huettel, 2012; Lee & Seo, 2016). Combining formal models of mentalizing with reinforcement learning is important to understand how social information is utilized by brain regions involved in mentalizing and strategic reasoning together with brain network supporting learning. A behavioral study (Lindström & Olsson, 2015) examining copying the behavior of others, showed that observers displayed an especially strong tendency to copy the observed behavior when they believed that the wrong decisions might be punished by a shock, as compared to when they thought their actions might be rewarded. Computational modeling showed that observers assigned the value to the other's actions, and used this to guide their own behavior. This shows how the tendency to copy others, combined with basic learning mechanisms, can generate and maintain behaviors. It is likely that brain regions involved in both non-social and social RL, such as the ventral striatum, contributes to these computational mechanisms.

When learning actions from others, it is important to learn from the "right" individual. Indeed, both human children and chimpanzees appear to use a "copy knowledgeable others" strategy (Kendal et al., 2015; Wood, Kendal, & Flynn, 2013), and adult humans rely more on social learning if demonstrators are described as skilled versus unskilled even if they perform equally well (Selbing & Olsson, 2017). Recent studies have described the neuro-computational mechanisms that might underlie such inferences. For example, learning about other individuals' preferences, used by an observer to infer their value as demonstrators, can be supported by a type of reinforcement learning that might be specific to social behavior: "inverse RL." In an experiment targeting this form of learning, Collette and colleagues (Collette, Pauli, Bossaerts, & O'Doherty, 2017) asked participants to watch the choices of demonstrators whom they were informed had either similar or dissimilar food preferences to themselves. Because only the demonstrator's actions, but never the decision outcome, were visible, participants could not base their own decisions on knowledge about the consequences of different choice options. Interestingly, instead of just imitating the demonstrators' actions, participants used inverse RL, meaning that they inferred the reward distribution for the actions of others solely through observing the actions implemented by another agent (Collette et al., 2017; Ng & Russell, 2000) allowing for learning from others with diametrically opposing preferences from the self. The imaging results revealed that the inverse RL updating signals were represented in the TPJ and the STS, key regions of the mentalizing network.

The study of mentalizing processes during social learning is key to our understanding of inter-individual communication, and how these interactions with others shape our attitudes and behavior. Recent efforts are making a head way using paradigms to study real-time interaction between participants (Liu et al., 2018; Schilbach, 2014; Stanley & Adolphs, 2013). In an experiment targeting the neural mechanisms underlying consensus decisions in social groups, a participant in the scanner interacted in real time with a group of participants (Suzuki, Adachi, Dunne, Bossaerts, & O'Doherty, 2015). The results demonstrated that people reached consensus through integrating own preferences with the preference of the majority, modulated by an estimate of how much each option was stuck to by the others in the group. Computational modeling showed that key social decision-making variables were encoded in different brain regions. Importantly, whereas own preferences were encoded in the vmPFC, the preferences of the majority were computed in the STS and TPJ, suggesting the involvement of mentalizing-related processes also in this kind of learning. In a related study, participants first rated how much they valued various consumer goods, and were then exposed to the preferences of several others. The results showed that participants updated their initial value judgments in a Bayesian fashion, computing both the subjective uncertainty of their initial beliefs and the reliability of the social information. Moreover, the dmPFC was found to track the degree of belief update. The authors argued that, analogous to how lower-level perceptual information is integrated, social information is integrated according to its reliability when judging value and confidence. These and other similar studies illustrate how computational and neural properties of social learning can be studied in a naturalistically dynamic social environment.

Concluding Remarks and Future Directions

In this chapter, we have surveyed research showing that both vicarious Pavlovian and vicarious instrumental learning share basic behavioral, computational, and neural principles with self-experienced learning. Importantly, we have described how these presumably domain-general learning principles are regulated by a range of social cognitive factors, which, to different degrees, imply the involvement of mentalizing processes computed in a distributed neural network, including the STS, TPJ, and the MPFC. Although few studies have directly manipulated mentalizing during social learning, a growing number of behavioral and neuroimaging studies have measured and modeled the computations underlying mentalizing during social learning. Taken together with an increasing emphasis on employing naturalistic situations, these approaches have moved the field forward in terms of mechanistic understanding aspects of social learning that might be both generalizable across situations, yet sensitive to situational factors. One important conclusion is that the behavior (together with the attributed mental states) of a demonstrator can serve as an unconditioned stimulus that reinforces the observer's behavior. The ensued learning is thus partly determined by the specifics appraisals and attributions of the demonstrator.

To address the lack of knowledge regarding the specific role of various forms of mentalizing in vicarious learning, future work should continue applying computational approaches, increase efforts to develop and validate measures of mentalizing processes, and create better ways of directly manipulating these processes. Research would benefit from experiments tracking the specific cost/reward structure associated with various mentalizing and empathizing processes in an online fashion. This would inform us about the malleability of these processes. Related to this, by enhancing our understanding of the links between mentalizing and learning processes, research would enable the possibility to study long-term effects of interventions to change mentalizing in order to enhance healthy and prosocial behavior. Finally, research on mentalizing and learning should aim to bridge between levels of analysis to understand how these processes jointly scale up from the individual to social networks, and to larger group constellations. This would also allow for examining the role in shaping social norms and other large scale social phenomena. We are convinced that the study of vicarious learning provides a well-suited experimental paradigm to link the study of function and mechanism across brain, behavior, and various social phenomena.

Acknowledgments We thank Tove Hensler for assistance in preparing the manuscript, and her and Joana Vieira for comments on an earlier version of the manuscript. This research was supported by Knut and Alice Wallenberg Foundation Grant KAW 2014.0237, and a Swedish Research Council Consolidator grant 2018-00877 to Andreas Olsson. Armita Golkar was supported by Swedish Research Council grant 2017-02210.

References

Adolphs, R. (2008). The social brain: Neural basis of social knowledge. *Annual Review of Psychology, 60*(1), 693–716. https://doi.org/10.1146/annurev.psych.60.110707.163514

Allsop, S. A., Wichmann, R., Mills, F., Burgos-Robles, A., Chang, C. J., Felix-Ortiz, A. C., … Tye, K. M. (2018). Corticoamygdala transfer of socially derived information gates observational learning. *Cell, 173*(6), 1329–1342. https://doi.org/10.1016/j.cell.2018.04.004

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*(4), 268–277. https://doi.org/10.1038/nrn1884

Apps, M. A. J., Rushworth, M. F. S., & Chang, S. W. C. (2016). The anterior cingulate gyrus and social cognition: Tracking the motivation of others. *Neuron, 90*(4), 692–707. https://doi.org/10.1016/J.NEURON.2016.04.018

Askew, C., & Field, A. P. (2007). Vicarious learning and the development of fears in childhood. *Behaviour Research and Therapy, 45*, 2616–2627. https://doi.org/10.1016/j.brat.2007.06.008

Bandura, A. (1978). Social learning theory of aggression. *Journal of Communication, 28*(3), 12–29. https://doi.org/10.1111/j.1460-2466.1978.tb01621.x

Batson, C. D., Early, S., & Salvarani, G. (1997). Perspective taking: Imagining how another feels versus imagining how you would feel. *Personality and Social Psychology Bulletin, 23*(7), 751–758. https://doi.org/10.1177/0146167297237008

Berger, S. M. (1962). Conditioning through vicarious instigation. *Psychological Review, 69*(5), 450–466. https://doi.org/10.1037/h0046466

Bickart, K. C., Dickerson, B. C., & Feldman Barrett, L. (2014). The amygdala as a hub in brain networks that support social life. *Neuropsychologia, 63*, 235–248. https://doi.org/10.1016/j.neuropsychologia.2014.08.013

Boyd, R., & Richerson, P. J. (2009). Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1533), 3281–3288. https://doi.org/10.1098/rstb.2009.0134

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences of the United States of America, 107*(32), 14431–14436. https://doi.org/10.1073/pnas.1003111107

Carlin, J. D., & Calder, A. J. (2013). The neural basis of eye gaze processing. *Current Opinion in Neurobiology, 23*(3), 450–455. https://doi.org/10.1016/j.conb.2012.11.014

Carter, R. M., Bowling, D. L., Reeck, C., & Huettel, S. A. (2012). A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science, 337*(6090), 109–111. https://doi.org/10.1126/science.1219681

Chang, D.-J., & Debiec, J. (2016). Neural correlates of the mother-to-infant social transmission of fear. *Journal of Neuroscience Research, 94*(6), 526–534. https://doi.org/10.1002/jnr.23739

Chang, S. W. C., Gariépy, J.-F., & Platt, M. L. (2013). Neuronal reference frames for social decisions in primate frontal cortex. *Nature Neuroscience, 16*(2), 243–250. https://doi.org/10.1038/nn.3287

Collette, S., Pauli, W. M., Bossaerts, P., & O'Doherty, J. (2017). Neural computations underlying inverse reinforcement learning in the human brain. *eLife, 6*, e29718. https://doi.org/10.7554/eLife.29718

Cook, M., & Mineka, S. (1989). Observational conditioning of fear to fear-relevant versus fear-irrelevant stimuli in rhesus monkeys. *Journal of Abnormal Psychology, 98*(4), 448–459. https://doi.org/10.1037/0021-843X.98.4.448

De Martino, B., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T., & Love, B. C. (2017). Social information is integrated into value and confidence judgments according to its reliability. *Journal of Neuroscienc, 37*(25), 6066–6074. https://doi.org/10.1523/JNEUROSCI.3880-16.2017

Debiec, J., & Olsson, A. (2017). Social fear learning: From animal models to human function. *Trends in Cognitive Sciences, 21*(7), 546–555. https://doi.org/10.1016/j.tics.2017.04.010

Debiec, J., & Sullivan, R. M. (2014). Intergenerational transmission of emotional trauma through amygdala-dependent mother-to-infant transfer of specific fear. *Proceedings of the National Academy of Sciences of the United States of America, 111*(33), 12222–12227. https://doi.org/10.1073/pnas.1316740111

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience, 24*(8), 1742–1752. https://doi.org/10.1162/jocn_a_00233

Eippert, F., Bingel, U., Schoell, E., Yacubian, J., & Buchel, C. (2008). Blockade of endogenous opioid neurotransmission enhances acquisition of conditioned fear in humans. *Journal of Neuroscience, 28*(21), 5465–5472. https://doi.org/10.1523/JNEUROSCI.5336-07.2008

Ekman, P. (1982). Methods for measuring facial action. In K. R. Scherer & P. Ekman (Eds.), *Handbook of methods in nonverbal behavior research* (pp. 45–135). New York, NY: Cambridge University Press.

Espinosa, L., Kleberg, J. L., Hofvander, B., Berggren, S., Bölte, S., & Olsson, A. (in press). Enhanced social learning of threat in adults with autism.

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology, 63*, 287–313. https://doi.org/10.1146/annurev-psych-120710-100449

Fullana, M. A., Harrison, B. J., Soriano-Mas, C., Vervliet, B., Cardoner, N., Àvila-Parcet, A., & Radua, J. (2016). Neural signatures of human fear conditioning: An updated and extended meta-analysis of fMRI studies. *Molecular Psychiatry, 21*(4), 500. https://doi.org/10.1038/mp.2015.88

Gerull, F. C., & Rapee, R. M. (2002). Mother knows best: Effects of maternal modelling on the acquisition of fear and avoidance behaviour in toddlers. *Behaviour Research and Therapy, 40*(3), 279–287. https://doi.org/10.1016/S0005-7967(01)00013-4

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences of the United States of America, 108*, 15647–15654. https://doi.org/10.1073/pnas.1014269108

Goldman, A. I. (1992). In defense of the simulation theory. *Mind & Language, 7*(1–2), 104–119. https://doi.org/10.1111/j.1468-0017.1992.tb00200.x

Golkar, A., Castro, V., & Olsson, A. (2015). Social learning of fear and safety is determined by the demonstrator's racial group. *Biology Letters, 11*(1), 20140817. https://doi.org/10.1098/rsbl.2014.0817

Golkar, A., Haaker, J., Selbing, I., & Olsson, A. (2016). Neural signals of vicarious extinction learning. *Social Cognitive and Affective Neuroscience, 11*(10), 1541–1549. https://doi.org/10.1093/scan/nsw068

Golkar, A., & Olsson, A. (2017). The interplay of social group biases in social threat learning. *Scientific Reports, 7*(1), 7685. https://doi.org/10.1038/s41598-017-07522-z

Golkar, A., Selbing, I., Flygare, O., Öhman, A., & Olsson, A. (2013). Other people as means to a safe end. *Psychological Science, 24*(11), 2182–2190. https://doi.org/10.1177/0956797613489890

Haaker, J., Golkar, A., Selbing, I., & Olsson, A. (2017). Assessment of social transmission of threats in humans using observational fear conditioning. *Nature Protocols, 12*, 1378. https://doi.org/10.1038/nprot.2017.027

Haaker, J., Molapour, T., & Olsson, A. (2016). Conditioned social dominance threat: Observation of others' social dominance biases threat learning. *Social Cognitive and Affective Neuroscience, 11*(10), 1627–1637. https://doi.org/10.1093/scan/nsw074

Haaker, J., Yi, J., Petrovic, P., & Olsson, A. (2017). Endogenous opioids regulate social threat learning in humans. *Nature Communications, 8*, 15495. https://doi.org/10.1038/ncomms15495

Hein, G., Engelmann, J. B., Vollberg, M. C., & Tobler, P. N. (2016). How learning shapes the empathic brain. *Proceedings of the National Academy of Sciences of the United States of America, 113*(1), 80–85. https://doi.org/10.1073/pnas.1514539112

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences, 6*(6), 242–247.

Heyes, C. (2012). What's social about social learning? *Journal of Comparative Psychology, 126*(2), 193–202. https://doi.org/10.1037/a0025180

Hill, C. A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J. P., & Ruff, C. C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience, 20*(8), 1142–1149. https://doi.org/10.1038/nn.4602

Hopwood, T. L., & Schutte, N. S. (2017). Psychological outcomes in reaction to media exposure to disasters and large-scale violence: A meta-analysis. *Psychology of Violence, 7*(2), 316–327. https://doi.org/10.1037/vio0000056

Hume, D. (1985). *A treatise of human nature*. London: Penguin Books. (reprinted; E. C. Mossner, Ed.).

Hygge, S., & Öhman, A. (1978). Modeling processes in the acquisition of fears: Vicarious electrodermal conditioning to fear-relevant stimuli. *Journal of Personality and Social Psychology, 36*(3), 271–279. https://doi.org/10.1037/0022-3514.36.3.271

Jeon, D., Kim, S., Chetana, M., Jo, D., Ruley, H. E., Lin, S.-Y., … Shin, H.-S. (2010). Observational fear learning involves affective pain system and Cav1.2 Ca2+ channels in ACC. *Nature Neuroscience, 13*(4), 482–488. https://doi.org/10.1038/nn.2504

Joiner, J., Piva, M., Turrin, C., & Chang, S. W. C. (2017). Social learning through prediction error in the brain. *NPJ Science of Learning, 2*(1), 8. https://doi.org/10.1038/s41539-017-0009-2

Jones, C. E., Riha, P. D., Gore, A. C., & Monfils, M.-H. (2014). Social transmission of Pavlovian fear: Fear-conditioning by-proxy in related female rats. *Animal Cognition, 17*(3), 827–834. https://doi.org/10.1007/s10071-013-0711-2

Kavaliers, M., Choleris, E., & Colwell, D. D. (2001). Learning from others to cope with biting flies: Social learning of fear-induced conditioned analgesia and active avoidance. *Behavioral Neuroscience, 115*(3), 661–674. https://doi.org/10.1037/0735-7044.115.3.661

Kendal, R., Hopper, L. M., Whiten, A., Brosnan, S. F., Lambeth, S. P., Schapiro, S. J., & Hoppitt, W. (2015). Chimpanzees copy dominant and knowledgeable individuals: Implications for cultural diversity. *Evolution and Human Behavior, 36*(1), 65–72. https://doi.org/10.1016/j.evolhumbehav.2014.09.002

Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., & Jones, P. L. (2018). Social learning strategies: Bridge-building between fields. *Trends in Cognitive Sciences, 22*(7), 651–665. https://doi.org/10.1016/j.tics.2018.04.003

Keysers, C., & Gazzola, V. (2014). Dissociating the ability and propensity for empathy. *Trends in Cognitive Sciences, 18*(4), 163–166. https://doi.org/10.1016/j.tics.2013.12.011

Kleberg, J. L., Selbing, I., Lundqvist, D., Hofvander, B., & Olsson, A. (2015). Spontaneous eye movements and trait empathy predict vicarious learning of fear. *International Journal of Psychophysiology, 98*(3), 577–583. https://doi.org/10.1016/j.ijpsycho.2015.04.001

Knapska, E., Mikosz, M., Werka, T., & Maren, S. (2010). Social modulation of learning in rats. *Learning & Memory, 17*(1), 35–42. https://doi.org/10.1101/lm.1670910

Kravetz, D. F. (1974). Heart rate as a minimal cue for the occurrence of vicarious classical conditioning. *Journal of Personality and Social Psychology, 29*(1), 125–131. https://doi.org/10.1037/h0035679

Lamm, C., Batson, C. D., & Decety, J. (2007). The neural substrate of human empathy: Effects of perspective-taking and cognitive appraisal. *Journal of Cognitive Neuroscience, 19*(1), 42–58. https://doi.org/10.1162/jocn.2007.19.1.42

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage, 54*(3), 2492–2502. https://doi.org/10.1016/j.neuroimage.2010.10.014

Lanzetta, J. T., & Englis, B. G. (1989). Expectations of cooperation and competition and their effects on observers' vicarious emotional responses. *Journal of Personality and Social Psychology, 56*(4), 543–554. https://doi.org/10.1037/0022-3514.56.4.543

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience, 23*(1), 155–184. https://doi.org/10.1146/annurev.neuro.23.1.155

Lee, D., & Seo, H. (2016). Neural basis of strategic decision making. *Trends in Neurosciences, 39*(1), 40–48. https://doi.org/10.1016/j.tins.2015.11.002

Lindström, B., Golkar, A., Jangard, S., Tobler, P. N., & Olsson, A. (2019). Social to instrumental transfer of fear in human decision-making. *Proceedings of the National Academy of Sciences of the United States of America, 116*, 4732. https://doi.org/10.1073/pnas.1810180116

Lindström, B., Haaker, J., & Olsson, A. (2018). A common neural network differentially mediates direct and social fear learning. *NeuroImage, 167*, 121–129. https://doi.org/10.1016/j.neuroimage.2017.11.039

Lindström, B., & Olsson, A. (2015). Mechanisms of social avoidance learning can explain the emergence of adaptive and arbitrary behavioral traditions in humans. *Journal of Experimental Psychology: General, 144*(3), 688–703. https://doi.org/10.1037/xge0000071

Liu, D., Liu, S., Liu, X., Zhang, C., Li, A., Jin, C., … Zhang, X. (2018). Interactive brain activity: Review and progress on EEG-based hyperscanning in social interactions. *Frontiers in Psychology, 9*, 1862. https://doi.org/10.3389/fpsyg.2018.01862

Lockwood, P. L., Apps, M. A. J., Roiser, J. P., & Viding, E. (2015). Encoding of vicarious reward prediction in anterior cingulate cortex and relationship with trait empathy. *The Journal of Neuroscience, 35*(40), 13720–13727. https://doi.org/10.1523/JNEUROSCI.1703-15.2015

Marshall, P. J., & Meltzoff, A. N. (2014). Neural mirroring mechanisms and imitation in human infants. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*, 20130620. https://doi.org/10.1098/rstb.2013.0620

Meffert, H., Brislin, S. J., White, S. F., & Blair, J. R. (2015). Prediction errors to emotional expressions: The roles of the amygdala in social referencing. *Social Cognitive and Affective Neuroscience, 10*(4), 537–544. https://doi.org/10.1093/scan/nsu085

Meyza, K. Z., Bartal, I. B.-A., Monfils, M. H., Panksepp, J. B., & Knapska, E. (2017). The roots of empathy: Through the lens of rodent models. *Neuroscience & Biobehavioral Reviews, 76*, 216–234. https://doi.org/10.1016/j.neubiorev.2016.10.028

Monfils, M. H., & Agee, L. A. (2019). Insights from social transmission of information in rodents. *Genes, Brain and Behavior, 18*(1), e12534. https://doi.org/10.1111/gbb.12534

Montague, P. R., Hyman, S. E., & Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature, 431*(7010), 760–767. https://doi.org/10.1038/nature03015

Morelli, S. A., Knutson, B., & Zaki, J. (2018). Neural sensitivity to personal and vicarious reward differentially relate to prosociality and well-being. *Social Cognitive and Affective Neuroscience, 13*(8), 831–839. https://doi.org/10.1093/scan/nsy056

Ng, A. Y., & Russell, S. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, Stanford, CA (pp. 663–670).

Ochsner, K. N., Ray, R. D., Cooper, J. C., Robertson, E. R., Chopra, S., Gabrieli, J. D. E., & Gross, J. J. (2004). For better or for worse: Neural systems supporting the cognitive down- and up-regulation of negative emotion. *NeuroImage, 23*(2), 483–499. https://doi.org/10.1016/j.neuroimage.2004.06.030

Olsson, A., Knapska, E., & Lindström, B. (in press). The neural systems of social learning.

Olsson, A., McMahon, K., Papenberg, G., Zaki, J., Bolger, N., & Ochsner, K. N. (2016). Vicarious fear learning depends on empathic appraisals and trait empathy. *Psychological Science, 27*(1), 25–33. https://doi.org/10.1177/0956797615604124

Olsson, A., Nearing, K. I., & Phelps, E. A. (2007). Learning fears by observing others: The neural systems of social fear transmission. *Social Cognitive and Affective Neuroscience, 2*(1), 3–11. https://doi.org/10.1093/scan/nsm005

Olsson, A., & Ochsner, K. N. (2008). The role of social cognition in emotion. *Trends in Cognitive Sciences, 12*(2), 65–71. https://doi.org/10.1016/j.tics.2007.11.010

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience, 10*(9), 1095–1102. https://doi.org/10.1038/nn1968

Olsson, A., & Spring, V. (2018). The vicarious brain: Integrating empathy and emotional learning. In K. Z. Meyza & E. Knapska (Eds.), *Neuronal correlates of empathy: From rodent to human*. New York, NY: Academic Press. ISBN: 9780128093481.

Pan, Y., Novembre, G., Song, B., Li, X., & Hu, Y. (2018). Interpersonal synchronization of inferior frontal cortices tracks social interactive learning of a song. *NeuroImage, 183*, 280–290. https://doi.org/10.1016/j.neuroimage.2018.08.005

Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2018). Similar neural responses predict friendship. *Nature Communications, 9*(1), 332. https://doi.org/10.1038/s41467-017-02722-7

Pärnamets, P., Espinosa, L., & Olsson, A. (2019). Physiological synchrony predicts observational threat learning in humans. *BioRxiv*. https://doi.org/10.1101/454819

Pavlov, I. P. (1927). *Conditional reflexes: An investigation of the physiological activity of the cerebral cortex*. Oxford, England: Oxford University Press.

Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron, 48*(2), 175–187. https://doi.org/10.1016/J.NEURON.2005.09.025

Rachman, S. (1977). The conditioning theory of fear acquisition: A critical examination. *Behaviour Research and Therapy, 15*, 375. https://doi.org/10.1016/0005-7967(77)90041-9

Repacholi, B. M., & Meltzoff, A. N. (2007). Emotional eavesdropping: Infants selectively respond to indirect emotional signals. *Child Development, 78*(2), 503–521. https://doi.org/10.1111/j.1467-8624.2007.01012.x

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*(1), 169–192. https://doi.org/10.1146/annurev.neuro.27.070203.144230

Rolls, E. T. (1999). *The brain and emotion*. New York, NY: Oxford University Press.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience, 15*, 549–562. https://doi.org/10.1038/nrn3776

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporoparietal junction. *Neuropsychologia, 43*(10), 1391–1399. https://doi.org/10.1016/j.neuropsychologia.2005.02.013

Schilbach, L. (2014). On the relationship of online and offline social cognition. *Frontiers in Human Neuroscience, 8*, 278. https://doi.org/10.3389/fnhum.2014.00278

Selbing, I., & Olsson, A. (2017). Beliefs about others' abilities alter learning from observation. *Scientific Reports, 7*(1), 16173. https://doi.org/10.1038/s41598-017-16307-3

Seymour, B., Singer, T., & Dolan, R. (2007). The neurobiology of punishment. *Nature Reviews Neuroscience, 8*(4), 300–311. https://doi.org/10.1038/nrn2119

Shu, J., Hassell, S., Weber, J., Ochsner, K. N., & Mobbs, D. (2017). The role of empathy in experiencing vicarious anxiety. *Journal of Experimental Psychology: General, 146*(8), 1164–1188. https://doi.org/10.1037/xge0000335

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science, 303*(5661), 1157–1162. https://doi.org/10.1126/science.1093535

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature, 439*(7075), 466–469. https://doi.org/10.1038/nature04271

Stanley, D. A., & Adolphs, R. (2013). Toward a neural basis for social behavior. *Neuron, 80*(3), 816–826. https://doi.org/10.1016/j.neuron.2013.10.038

Sullivan, R. M., Landers, M., Yeaman, B., & Wilson, D. A. (2000). Good memories of bad events in infancy. *Nature, 407*(6800), 38–39. https://doi.org/10.1038/35024156

Suzuki, S., Adachi, R., Dunne, S., Bossaerts, P., & O'Doherty, J. P. (2015). Neural mechanisms underlying human consensus decision-making. *Neuron, 86*(2), 591–602. https://doi.org/10.1016/j.neuron.2015.03.019

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences, 22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005

Tomasello, M. (2011). Human culture in evolutionary perspective. In M. Gelfand, C. Chiu, & Y. Hong (Eds.), *Advances in culture and psychology* (Vol. 1, pp. 5–51). New York, NY: Oxford University Press.

Williams, A., & Conway, C. (2019). Empathy does not amplify vicarious threat learning. *PsyArXiv*.

Wood, L. A., Kendal, R. L., & Flynn, E. G. (2013). Whom do children copy? Model-based biases in social learning. *Developmental Review, 33*(4), 341–356. https://doi.org/10.1016/j.dr.2013.08.002

Zaki, J. (2014). Empathy: A motivated account. *Psychological Bulletin, 140*(6), 1608–1647. https://doi.org/10.1037/a0037679

Zaki, J., & Ochsner, K. (2012). The neuroscience of empathy: Progress, pitfalls and promise. *Nature Neuroscience, 15*(5), 675–680. https://doi.org/10.1038/nn.3085

An Examination of Accurate Versus "Biased" Mentalizing in Moral and Economic Decision-Making

BoKyung Park, Minjae Kim, and Liane Young

Value-based decision-making, especially in social contexts, depends critically on the ability to think about agents' mental states, i.e., mentalizing or Theory of Mind (ToM). Although the involvement of mentalizing in different decision-making processes, such as moral judgment and economic exchange, is generally acknowledged, whether mentalizing leads to optimal or suboptimal decisions is a relatively open question. Accurate mentalizing leads to optimal decisions that maximize immediate or future benefits, including learning from others, defeating others, evaluating others, and predicting others. Yet, mentalizing is also vulnerable to "bias"; mentalizing is affected by a number of ostensibly irrelevant factors, including the identity and group status of the interacting agents, the mentalizer's own beliefs and values, and other contextual factors. We suggest that, in these cases, mentalizing can lead to suboptimal decisions. In the last section of this chapter, we revisit cases of mentalizing that appear to be biased, taking ingroup bias as a case study, and we suggest that a subset of these cases may be compatible with rational Bayesian reasoning.

Thus, in this chapter, we review cases in which mentalizing supports both optimal and suboptimal value-based decisions, in the domains of moral judgment and economic exchange. We will also examine how seemingly biased mentalizing and subsequent suboptimal decisions may in fact arise from a rational procedure.

Mentalizing Network

Before we discuss the role of mentalizing in moral and economic decision-making, we briefly summarize research on the network of brain regions that support mentalizing, also known as the ToM network. Decades of work, using functional magnetic

B. Park (✉) · M. Kim (✉) · L. Young
Department of Psychology, Boston College, Chestnut Hill, MA, USA
e-mail: parkanj@bc.edu; minjae.kim@bc.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_27

resonance imaging (fMRI) and event-related potential (ERP) methods, points to several key nodes: the right temporo-parietal junction (rTPJ; often labeled as posterior superior temporal cortex, inferior parietal lobule, or Brodmann area 39), left temporo-parietal junction (lTPJ), posterior superior temporal sulcus (pSTS), dorsomedial prefrontal cortex (dmPFC), and precuneus (Decety & Cacioppo, 2012; Saxe, 2009; Saxe, Carey, & Kanwisher, 2004; Saxe & Kanwisher, 2003; Saxe & Powell, 2006; Saxe & Wexler, 2005; Saxe, Xiao, Kovacs, Perrett, & Kanwisher, 2004). Recent empirical and theoretical evidence suggests that these regions support mentalizing by, to some extent, encoding social prediction error, i.e., they respond preferentially to unexpected agent behaviors (Koster-Hale & Saxe, 2013). Other work reveals that other sub-regions of the medial prefrontal cortex (mPFC) including ventromedial prefrontal cortex (vmPFC), anterior cingulate cortex (ACC), and adjacent paracingulate cortex are also recruited for mentalizing (Amodio & Frith, 2006; Frith & Frith, 2006; Krueger, Grafman, & McCabe, 2008; Lombardo et al., 2009; Walter et al., 2004). Of these regions, the MPFC and bilateral TPJ emerged as consistently activated across ToM tasks in a massive activation likelihood estimation (ALE) meta-analyses of 144 datasets (3150 participants) (Molenberghs, Johnson, Henry, & Mattingley, 2016).

Accurate Mentalizing Leads to Optimal Decisions

Evidence demonstrates that accurate mentalizing can result in immediate rewards, such as earning money, or more distant rewards, such as identifying future cooperators versus competitors.

First, given the primacy of moral signals in impression updating compared to other trait information (Brambilla, Carraro, Castelli, & Sacchi, 2019; Goodwin, 2015), moral judgment—evaluating whether an agent's behavior is right or wrong—is crucial for identifying potential friend versus foe, and maximizing future social benefits. A large body of previous research has identified the key role of mentalizing regions, and specifically the rTPJ, in the formation and revision of moral judgments (e.g., Decety & Cacioppo, 2012; Young, Cushman, Hauser, & Saxe, 2007). Specifically, rTPJ activity is consistently recruited for intent-based moral judgments, including: forgiving accidents (innocent intent), condemning failed attempts to harm (malicious intent) (Young, Nichols, & Saxe, 2010; Young & Saxe, 2009), and even withholding praise for unintentionally helpful behaviors (Young, Scholz, & Saxe, 2011). Spatial patterns of activity in rTPJ discriminate between intentional and accidental harms, and also correlate with moral judgments, though this pattern discrimination is absent in high-functioning adults with autism (Koster-Hale, Saxe, Dungan, & Young, 2013). Moreover, mentalizing supports the integration of mitigating intent information even for extreme harms (e.g., killing one's wife to relieve her suffering); reduced punishment was associated with increased rTPJ activity (Yamada et al., 2012). Other work has suggested that forgiving accidents may involve suppressing emotional responses to negative outcomes, indexed by greater coupling between the mentalizing network activity and amygdala activity in response to unintended harms (Treadway et al., 2014).

Convergent transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS) studies have shown that modulating rTPJ activity leads to systematically different moral judgments, establishing a causal role for the rTPJ in mentalizing for moral judgment. Disrupting rTPJ activity using TMS leads to more outcome-based moral judgments (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010), whereas increasing the excitability of the rTPJ with tDCS leads to more intent-based moral judgments (Sellaro et al., 2015).

Developmental work reveals that young children aged 3–4 years, who lack mature mentalizing capacities, fail to incorporate the intention information and make more outcome-based moral judgments as well (Killen, Mulvey, Richardson, Jampol, & Woodward, 2011; see also Cushman, Sheketoff, Wharton, & Carey, 2013). The ability to integrate mental state information with other task-relevant information for moral judgment is supported by developing neural circuitry, including the rTPJ, for mentalizing (Gweon, Dodell-Feder, Bedny, & Saxe, 2012; Richardson, Lisandrelli, Riobueno-Naylor, & Saxe, 2018).[1] Thus, accurately inferring the mental states of others helps people to appropriately assign blame and praise to agents, contributing to future cooperative relationships.

In addition to its role in identifying social partners with whom it is a good idea to cooperate, mentalizing can also be a useful tool in competitive contexts—situations in which we have to figure out what other agents are thinking in order to predict and outsmart them (Singer & Fehr, 2005). In one study, when participants were asked to play a game against strategic human partners, greater engagement of MPFC was associated with better game performance (Coricelli & Nagel, 2009). RTPJ was also recruited when participants defected against their partners to earn greater profits (Bhatt, Lohrenz, Camerer, & Montague, 2010), suggesting a role for mental state inference in strategizing against other agentic opponents.

Meanwhile, when people engage in repeated interactions with the same partner, they must track their partner’s actions and select optimal interaction strategies—processes that also depend on mentalizing (Lee, 2006). Ample evidence shows that the mentalizing network supports the processes by which people update their representations of others’ personality traits when their behaviors change meaningfully over time (Baron, Gobbini, Engell, & Todorov, 2011; Mende-Siedlecki, Baron, & Todorov, 2013; Mende-Siedlecki, Cai, & Todorov, 2013; Thornton & Mitchell, 2018; see also Mende-Siedlecki & Todorov, 2016).

Additional research has investigated more complex contexts in which people had to interactively revise their own behavior in response to the behavior of other agents who could impact their outcomes (Hampton, Bossaerts, & O’Doherty, 2008). Participants were paired with a partner and took turns playing as the “employee,”

[1] Other work has focused not on the inference of mental states but the inference of moral traits, including generosity. In one study, trait generosity (i.e., proportion of money an agent offered) was encoded separately from total reward provided by the agent, in the rTPJ (Hackel et al., 2015). Partner choice decisions relied primarily on trait generosity.

who could either work or shirk, or the "employer," who could either inspect or not inspect ("inspection game"). The employee earned money when she shirked and the employer did not inspect, or when she worked and the employer did inspect. In contrast, the employer earned a reward when she did not inspect and the employee worked, or when she inspected and the employee shirked. Thus, accurate prediction of the partner's next action should be based on the history of her choices, as well as the fact that the participant's own action can in turn modify the partner's behavior. The researchers found that activity in the STS tracked updating of the partner's strategy based on this computation. Moreover, MPFC and ventral striatum were conjointly recruited and highly correlated with STS; these two regions encoded different components of expected reward from the interactions. These findings suggest that this interactive network supports the revision of decisions based on action valuation and mentalizing. By coordinating these different networks, this process potentially generates more fine-grained representations of opponents' future actions. A recent study using a similar inspection game paradigm probed the causal impact of mentalizing on interactive updating using TMS (Hill et al., 2017). When participants whose rTPJ was disrupted by TMS played as the employee, they failed to consider the causal link between their own actions and the employer's future behavior.[2]

Accurately forming and revising representations of another agent's mental states is also important in cases where that agent has privileged access to useful information. One example is when a decision-maker considers advice from others ("advisors"). Decisions to follow the advice or not depend on the advisor's mental states, including her intention and expertise (Harvey & Fischer, 1997; Klucharev, Smidts, & Fernández, 2008; Schilbach, Eickhoff, Schultze, Mojzisch, & Vogeley, 2013; Van Swol, 2009; Yaniv & Kleinberger, 2000). In one study, participants were asked to choose one of two fixed options that would likely return greater scores (Behrens, Hunt, Woolrich, & Rushworth, 2008). An ostensible advisor gave advice to participants across trials, but the advisor's goal was to ensure that participants score within a certain limited range, not to maximize the score. Thus, participants needed to keep track of two different elements based on each outcome: the predicted scores associated with each of the two choice options, and the current intention of the advisor as a function of their current score. While participants' reward computations (in a reinforcement learning model) were tracked by reward-processing regions (ventral striatum and vmPFC), critically, computations of the advisor's intention were tracked by nodes in the mentalizing network (dmPFC and rTPJ). The two information sources were combined in vmPFC. Thus, to maximize personal value, participants recruited different brain regions, including the mentalizing network, to update representations of the reward and the advisor. In another study, participants made

[2] However, another body of research failed to find that repeated interaction with partners necessarily involves mentalizing. These studies focused on the involvement of the reward-processing circuitry, or interpreted activity of regions that constitute both mentalizing and reward valuation networks (e.g., MPFC) in the light of reward computation. See Delgado, Frank, and Phelps (2005), Fareri et al. (2015), Izuma, Saito, and Sadato (2008), King-Casas et al. (2005), and Phan, Sripada, Angstadt, and McCabe (2010), as examples.

predictions about whether an asset value would increase or decrease on some trials (Boorman, O’Doherty, Adolphs, & Rangel, 2013). More importantly, on other trials, there were advisors who made the predictions about asset value, and participants had to bet for or against the advisor before the prediction was revealed; they earned a bonus for correct bets. Thus, tracking the expertise of the advisors was important. Over the course of the game, participants could use the feedback to form and update beliefs about the expertise of the advisors. When participants learned about the expertise of the advisor, based on whether the advisor’s choice agreed with their own prediction or not, they recruited the mentalizing network, including rTPJ and dmPFC. Thus, mentalizing regions may have supported participants’ capacity to generate accurate representations of the advisor’s expertise.

Similarly, people can reflect on others’ choices to infer relative values of available options to maximize their own rewards. An interesting case of this is how people navigate fluctuating stock markets. A certain stock price rises, and people have to infer what other traders think. Does this pattern reflect the stock’s real value, or arbitrary noise from other sources? Bruguier, Quartz, and Bossaerts (2010) found that mentalizing can support optimal decisions in these contexts, especially when other traders in the market were known to have better access to critical information (“insiders”). When there were insiders who knew the specific dividends of stocks in the market, price changes of stocks in these markets were a diagnostic tool to estimate their dividends. Participants were informed whether there were insiders in the current market or not, and chose the number of shares of different types of stocks. Of key relevance, mentalizing regions, including the medial paracingulate cortex, were recruited more when insiders were present than absent. Participants’ mentalizing ability was independently measured in two separate tasks: where they (1) predicted agentic movement of shapes and (2) inferred mental states from eye gazes. Participants who were better at mentalizing performed better in forecasting market trades, supporting the argument that accurate mentalizing for inferring insider strategies helps participants navigate the financial market more successfully (Bruguier et al., 2010).

To conclude this section, we reviewed evidence that mentalizing is critical for evaluating moral agents, including harmful agents, for predicting how competitors will behave, and for learning from those who have special information about a shared environment (e.g., stock markets). Thus, mentalizing can allow decision-makers to maximize profits in direct economic interactions, and to identify future cooperative partners through third-party observations. While mentalizing is essential for social decision-making, mentalizing can also go astray, as we will see in the next section.

Inaccurate Mentalizing Leads to Suboptimal Decisions

Although mentalizing supports successful social interaction, mentalizing is also susceptible to influence by factors that may be irrelevant to the decision at hand, including the moral character or group status of the target. Consequently, people may assign more or less blame than is warranted, leading to inaccurate identification of future friend or foe.

Prior research has manipulated participants' prior moral impressions of targets to investigate the impact on mentalizing. In one study, participants first interacted with fair and unfair agents; they then read vignettes describing good and bad actions presented as performed by the agents (Kliemann, Young, Scholz, & Saxe, 2008). Participants showed reduced rTPJ activity when a previously *fair* agent caused a negative outcome, compared to when a previously *unfair* agent caused a negative outcome. Furthermore, fair agents were judged as less blameworthy for causing negative outcomes, and these actions were judged as less intentional, compared the same negative outcomes caused by unfair agents. Together, these behavioral and neural patterns suggest that participants disengaged from intent attribution for previously fair agents. Consistent with this idea, a recent behavioral study revealed that when people initially had optimistic impressions of a financial advisor's expertise, they preferentially incorporated positive information about the advisor's accuracy and took the advisor's advice more than they should have given actual feedback (Leong & Zaki, 2017). Critically, when participants' initial impressions were directly manipulated to be more well-calibrated, the optimism bias went away. These findings suggest that when people make initially optimistic judgments about experts, they preferentially discount new evidence that is inconsistent with these judgments.

Another body of research suggests that salient negative outcomes of an agent's behavior can distort mentalizing processes. In one set of studies, participants were presented with vignettes describing a CEO causing environmental damage as a side effect of a new business policy (Knobe, 2003). Importantly, the CEO stated explicitly that he did not intend to cause environmental harm; however, participants who treated the environment as "sacred" perceived the harm as more intentional compared to participants who did not (Ditto, Pizarro, & Tannenbaum, 2009; Tannenbaum, Ditto, & Pizarro, 2008). Thus, morally unacceptable outcomes might lead participants to overestimate harmful intent. In broadly consistent studies, people blame agents who benefit from uncontrollable negative events, i.e., agents who bet on natural disasters, or agents who are forced to harm an enemy (Inbar, Pizarro, & Cushman, 2012; Woolfolk, Doris, & Darley, 2006; see also Pizarro, Uhlmann, & Bloom, 2003; Pizarro, Uhlmann, & Salovey, 2003).

Similar findings have emerged when participants themselves are actually impacted by bad or good behaviors. In one study, participants responded to offers from a partner in the ultimatum game, who either was forced to make an unfair offer or could choose between fair and unfair distributions (Güroğlu, van den Bos, Rombouts, & Crone, 2010). The researchers found that participants engaged in greater mentalizing, as indexed by greater rTPJ responses, when they rejected forced unfair offers compared to intended unfair offers. This finding suggests that people might justify their blame of faultless others by over-attributing harmful intent, while no mentalizing effort was required to reject the unambiguously intentional unfair offers.

Convergent evidence comes from studies examining the influence of group membership on mentalizing. Specifically, research has found that participants discount ingroup members’ negative behaviors (e.g., taking money from the participants; heckling a speaker during a talk) and thus fail to negatively update their impressions. In recent work, overcoming this bias to negatively evaluate a close friend (Park, Fareri, Delgado, & Young, 2020) or an ingroup member (Hughes, Zaki, & Ambady, 2017) was accompanied by recruitment of brain regions associated with mentalizing, including bilateral TPJ and ACC. A similar study examined this effect behaviorally. When participants were presented with an outgroup member’s negative behavior first, and intention information later, they increased blame for intentional harms to a greater degree than they reduced blame for unintentional harms. However, for ingroup members, participants used exacerbating and mitigating intent information symmetrically to assign blame (Monroe & Malle, 2019), suggesting that participants might engage in mentalizing more readily when they encounter the negative behavior of outgroup members versus ingroup members. Thus, group membership across diverse contexts influences when and how people engage in mentalizing, leading to occasionally inaccurate moral judgments.

The evidence reviewed thus far shows that people can disengage from mentalizing about targets for whom they have positive prior impressions, resulting in mitigated blame and reduced impression updating. However, another body of work indicates that *greater* mentalizing can also facilitate forgiveness and cooperation (Hare, Camerer, Knoepfle, O’Doherty, & Rangel, 2010; Krueger et al., 2007; Strang, Utikal, Fischbacher, Weber, & Falk, 2014; Will, Crone, & Güroğlu, 2015). Indeed, some studies found that greater mentalizing for ingroup members was associated with blame mitigation. Specifically, in one study, participants had the opportunity to punish ingroup and outgroup members who defected against another person in the prisoner’s dilemma game (Baumgartner, Götte, Gügler, & Fehr, 2012). When participants were presented with an ingroup defector, they showed increased activity in dmPFC and bilateral TPJ, reflecting an attempt to infer the intentions behind the defection. Moreover, increased connectivity among the nodes of the mentalizing network was associated with weaker punishment of ingroup members. Furthermore, disrupting rTPJ activity using TMS reduced forgiveness for ingroup members (Baumgartner, Schiller, Rieskamp, Gianotti, & Knoch, 2014). Another study found that the greater dmPFC activity participants showed when they played the prisoner’s dilemma game with ingroup compared to outgroup members, the more likely they were to cooperate with ingroup than outgroup members in the game (Rilling, Dagenais, Goldsmith, Glenn, & Pagnoni, 2008).

Thus, the act of mentalizing can lead to seemingly opposite consequences: exacerbating and mitigating blame (or decreasing and increasing cooperation). We will revisit this puzzle in the final section, where we explore whether these processes reflect rational versus motivated cognition. But, in either case, mentalizing serves the same purpose of preserving pre-existing impressions. For now, we note that mentalizing is vulnerable to the influence of irrelevant factors, which can lead to biased judgments and perhaps inaccurate action predictions.

Finally, inaccurate, biased mentalizing can also result in concrete financial losses. People often rely on others’ mental states to infer potential reward from future decisions, such as seeing other customers’ response to their food in a restaurant. Depending on the accuracy of the mental state inference, the value represented in one's mind may not reflect the real intrinsic value. A group of researchers tested this possibility in a paradigm that extended the target of mental state inference to a whole group of agents. Participants viewed experimental asset prices, some of which were inflated beyond their intrinsic value by crowds in the market, i.e., financial bubbles (De Martino, O'Doherty, Ray, Bossaerts, & Camerer, 2013). The researchers found that, compared to the non-bubble markets, in the bubble markets where participants had to infer intentions of other traders, the computed values of participants' current possession—reflecting the inflated value of their assets—were parametrically tracked by increased dmPFC activity as well as vmPFC activity. Moreover, there was greater functional coupling between dmPFC and vmPFC in the bubble market, and greater vmPFC activity was ultimately associated with greater likelihood of following the crowd in bubble markets. This pattern suggests that the computed intentions of other traders, reflected in dmPFC, were projected to vmPFC, a region associated with reward computation, perhaps leading participants to overestimate the role of intent in the rise of prices. Consequently, these participants purchased assets at high prices and ultimately earned less. Thus, observers who engage in excessive mentalizing for crowds may follow suboptimal trends and incur a financial loss.

Social interaction requires mentalizing; yet, as we have reviewed in this section, mentalizing is vulnerable to bias and can lead to suboptimal decisions. Prior moral impressions, which may be built through direct feedback, or implicitly signaled by group membership, can bias mentalizing, increasing the possibility of inaccurate mental state inferences. However, as we will explore in the final section, these biases may not reflect truly "irrational" processes. Although the resultant decisions, such as favorable judgments about ingroup members (and the discounting of negative information about ingroup members), may appear biased, the underlying *processes* may nevertheless be rational.[3] This idea may be the key to explaining why people sometimes engage in greater mentalizing, and other times less mentalizing, in order to protect positive impressions of ingroup members (and negative impressions of immoral agents). In the final section, we will discuss this puzzle and explore the possibility that seemingly biased social and moral judgments may actually reflect rational decision-making.

Motivated Mentalizing or Rational Updating?

Our prior knowledge of a person influences how we evaluate their behavior. Consider a close friend who you know to be trustworthy. One day, you see her take a quarter from a tip jar. Would you then judge her to be an untrustworthy person? Or—given

[3] Here we focus on procedural rationality, which may produce either accurate or inaccurate judgment. By contrast, see Cushman, 2020, for a theoretical account of how people ultimately benefit from rationalization.

your prior knowledge of her trustworthiness—would you consider this observation a noisy data point, and reattribute her behavior to situational factors? For instance: perhaps she was trying to make change for a dollar. By contrast, seeing a stranger take a quarter from a tip jar often leads to the inference that they are untrustworthy.

This asymmetry in our trait evaluations of friend versus stranger appears to be an instance of the well-known bias to positively evaluate close others or ingroup members. A key proposal of this chapter, however, is that the asymmetry can be accounted for by differences in the strength of prior knowledge. In the case of the stranger, we have no prior knowledge of their trustworthiness, so a single bad behavior is highly diagnostic of their character. But in the case of our friend, we have ample prior knowledge of her trustworthiness, so entirely revising our impression of her based on a single action may not be optimal. The confusing feature here is that strong prior knowledge of close others often co-occurs with factors that typically contribute to motivated decision-making, such as congenial affect, a long relationship history, and attachment. It is likely the case that both prior knowledge and socio-affective factors contribute to reduced belief updating in response to negative feedback; the relative contributions of these factors across contexts is a difficult but important empirical question. Here we highlight cases of seemingly motivated judgments that may instead be compatible with a rational updating process.

Bayesian updating provides a normative framework for how beliefs about others should be updated when new information is acquired. Bayes’ rule holds that the probability of a belief being true given new evidence—e.g., *P*(*my friend is trustworthy*|*she stole a quarter*)—is equal to the likelihood of the evidence being acquired given the prior belief, *P*(*she stole a quarter*|*she is trustworthy*), multiplied by the probability of the prior belief being true before receiving the new evidence, *P*(*she is trustworthy*), scaled by the probability of the new evidence being acquired, *P*(*she stole a quarter*). This process factors the strength of the prior belief into updating; it follows that new information that contradicts strong prior beliefs may be discounted. While Bayesian updating does not necessarily guarantee accurate mental state inference, it confers *procedural* rationality on the inference process (Hahn & Harris, 2014) and serves as a normative criterion for assessing deviations from rational belief updating (Hackel & Amodio, 2018). Why adopt Bayesian processing in particular as a criterion for rationality? According to a set of epistemological accounts called the “Dutch Book” arguments, when an agent possesses degrees of belief that violate the axioms of probability theory, they are vulnerable to logically ensured losses when acting on their beliefs (e.g., accepting a wager that will lead to a sure loss, regardless of outcome), and to internally inconsistent evaluations (see Hájek, 2008 for a review). By this account, adhering to the axioms of probability theory can protect us from holding beliefs that are logically guaranteed to be false, and which would impair utility maximization.

How can a Bayesian framework be used to understand the robustness of prior beliefs to contradictory evidence, especially in the case of moral updating? A theoretical account suggests that people can generate ad hoc auxiliary hypotheses to explain away evidence that contradicts prior beliefs, and that this process is Bayesian-rational (Gershman, 2019; see Lakatos, 1976 for discussion on the role of auxiliaries in science). This process adheres to probability theory: auxiliary hypotheses are more likely to be invoked when they are highly consistent with the new information, and when the prior belief has a relatively high probability. For instance, given your strong prior belief in the trustworthiness of your coin-taking friend, you may generate the auxiliary hypothesis that your friend was making change for a dollar. That is, the unexpected event is attributed to a situational cause, instead of a dispositional cause. While the tendency to invoke situational explanations for close others or ingroup members has been described as a cognitive bias, situational attributions can be procedurally rational if warranted by the strength of prior beliefs. Additionally, to return to the epistemological arguments for Bayesian rationality, invoking an auxiliary hypothesis in a graded manner allows the observer to retain a coherent set of beliefs that takes new evidence into account.[4]

How can we discern whether a case of reduced belief updating is the result of Bayesian-rational updating over strong priors, rather than non-rational discounting of contradictory evidence? Our novel proposal is that the rational route to belief preservation will recruit more mentalizing activity than the non-rational route. When a Bayesian observer is faced with new, meaningful information that is inconsistent with their prior evaluations, they can account for the discrepancy by updating their prior beliefs, or by generating an auxiliary hypothesis to explain away the information. We speculate that, at least in the domain of moral judgment and character evaluation, both of these processes will recruit the mentalizing network, in particular, rTPJ, given its role in supporting mental state-based moral judgment. Thus an association between increased rTPJ activity and increased updating may suggest Bayesian updating of prior beliefs, and an association between increased rTPJ activity and reduced updating may suggest the generation of auxiliary hypotheses. An association between *decreased* rTPJ activity and reduced updating, however, may suggest motivated discounting of new evidence.

We now apply this logic to several studies discussed above. Recall that Baumgartner et al. (2012) found increased mentalizing network activity in response to ingroup vs. outgroup defectors, and that greater connectivity in this network was associated with forgiveness of ingroup members. Increased mentalizing activity in this case can be reinterpreted as supporting the generation of auxiliary hypotheses that are consistent with strong positive beliefs about the ingroup. For example, perhaps the ingroup member did not intend to defect, or had a good reason to do so.

Turning to cases of motivated discounting, Kliemann et al. (2008) had found that, when a previously fair (vs. unfair) social partner was described as performing a harmful action, participants judged the action to be less intentional, and this judgment was associated with reduced rTPJ activity. If participants were taking a Bayesian route to belief maintenance, they would have engaged in more

[4] We also note that procedural rationality is orthogonal to the source of the prior belief: both priors that are evidence-based and priors that are derived largely from socio-affective value (e.g., positive beliefs about the ingroup in minimal group contexts) can undergo Bayesian processing.

mentalizing for fair partners, in order to explain away the evaluatively inconsistent information. We hypothesize that participants took a motivated route instead: they may have opted out of explaining the discrepancy by disengaging from mentalizing about fair partners, resulting in decreased inferences of harmful intent. The function of this selective disengagement may be to preserve a historically cooperative relationship. Further, disengagement from mentalizing can be seen for ingroup members as well. Generally, group membership may serve as a proxy for moral character, such that in the absence of direct evidence, ingroup members are viewed as good moral agents. Hughes et al. (2017) found that decreased rTPJ activity was associated with reduced impression updating in response to negative feedback for ingroup members, consistent with what the researchers termed "an effortless bias" account. In this case, participants who disengaged from mentalizing were able to maintain desirable beliefs about ingroup members, by failing to incorporate evidence that would have led to a negative character inference. These studies suggest that, in the face of evidence that affords disfavorable trait inferences about ingroup members or previously moral agents, people may opt out of rational updating by mentalizing less about these agents altogether. Future work should examine the role of decreased mentalizing in other contexts, such as economic games, in which people are resistant to updating in response to feedback about moral or ingroup targets (see Evans, Fleming, Dolan, & Averbeck, 2011; Fareri, Chang, & Delgado, 2015; Hackel, Doll, & Amodio, 2015).

Maintaining beliefs by discounting new information is not rational in the Bayesian sense, but it may be *adaptively* beneficial, in that it can increase social fitness and affective well-being. Specifically, it can be beneficial to maintain relationships with potential cooperative partners. For example, a group of researchers found that participants trusted their friends more than strangers in the trust game, even though reciprocation rates were equal for friend and stranger (Fareri et al., 2015). Neural and computational evidence indicated that trust decisions were driven by a striatal reward response to reciprocation from close friends. There can thus be affective benefits to non-rational processing of feedback about close others. Moreover, in recent research, we found that individuals who were more resistant to negatively updating their evaluations about a friend also reported having more friends in real life (Park & Young, 2020). These are cases in which reduced belief updating leads to inaccurate predictions and financially suboptimal decisions, but may ultimately maximize the affective and social benefits of interacting with and maintaining close friends.

Within a given context, individuals may vary in whether they take a procedurally rational or irrational path to belief maintenance. For example, one study examined the public's impressions of Bill Clinton 8 months before and 3 days after the Lewinsky story broke (Fischle, 2000). Respondents were interviewed on various aspects of the scandal, including the credibility and importance of the allegations, and attitudes towards the president's resignation. The study found that perceived importance of the scandal increased support for resignation by 57% for Clinton detractors, but only by 19% for Clinton supporters. The author argued that a Bayesian framework cannot account for such moderated effects, while a motivated reasoning process can capture affect-dependent weighting of factors like perceived importance. While this argument holds for those supporters who thought the allegations were important but did not support resignation, there were also supporters who exhibited—per our interpretation—the Bayesian response. In particular, this study also found that supporters were more likely than detractors to view the scandal as a conspiracy, and this reduced supporters' certainty of impropriety and their endorsement of resignation. Given supporters' robust prior beliefs about Clinton, this set of respondents may have generated the auxiliary hypothesis that the scandal was a conspiracy planted by the president's opponents. More generally, in studies that find motivation-derived evaluations, there may be individual differences in whether participants take a Bayesian route to belief maintenance, or deviate from Bayesian reasoning in order to maintain prior beliefs.

Comparing participants' behavioral belief updates with predictions from a Bayesian model of inference can reveal the contexts in which people engage in probabilistic belief updating. One recent study examined how people learn factual political statements based on noisy feedback from a computer, and found that participants closely followed Bayesian updating, but not perfectly (Hill, 2017). Specifically, participants evaluated the same factual statement across multiple rounds; in some rounds, the computer signaled, with 75% accuracy, whether the statement was true or not. Comparing participants' initial responses (prior beliefs) with their final responses (posterior beliefs), the author found that when the signal was consistent with their prior beliefs, participants did not deviate from what was expected by a Bayesian model. When the signal was inconsistent with their prior beliefs, however, participants updated less than expected by the Bayesian model, suggesting motivational influences. Importantly, a growing body of work has used a computational approach to investigate how people update their evaluations of others, such as when making repeated judgments of whether advisors are trustworthy and accurate. Some studies have found that observers are biased towards learning from evaluatively consistent information (e.g., Fareri, Chang, & Delgado, 2012; Leong & Zaki, 2017); others have found that participants derive inferences in a Bayesian-rational manner (Behrens et al., 2008; Cao, Kleiman-Weiner, & Banaji, 2019; Diaconescu et al., 2014; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015; see Hackel & Amodio, 2018 for a review of the literature). These findings highlight the utility of Bayesian models for systematically investigating when people adhere to, and when they deviate from, procedurally rational updating when learning about others' traits.

Neuroimaging studies, combined with a computational modeling approach, will provide an important window into the link between mentalizing network activity and motivated vs. Bayesian decision-making. Given that the rTPJ has been found to be engaged for probabilistic belief updating (Mengotii, Dombert, Fink, & Vossel, 2017), future research should combine fMRI and computational methods to further explore how the rTPJ may support Bayesian reasoning. Additionally, if there is a role for the rTPJ in Bayesian reasoning, it may be context-dependent. When the observer acquires evidence that does not warrant revision of a strong prior belief, the rTPJ may support the generation of auxiliary hypotheses (e.g., blame-mitigating mental states, appeal to situational factors); this may be the process underlying reduced punishment for ingroup defectors (Baumgartner et al., 2012). When the new evidence does warrant belief updating, the rTPJ may support revision of the strong prior belief (e.g., by attributing harmful intent). However, if the observer is motivated to maintain desired prior beliefs in the face of exceedingly strong contradictory evidence, their departure from Bayesian updating may be indexed by decreased rTPJ activity; this may underlie reduced negative updates for ingroup members (Hughes et al., 2017). Further work will be needed to characterize the conditions under which observers who have strong prior beliefs about targets will mentalize about them *less* upon receiving contradictory evidence—therefore opting out of drawing any inferences that would prompt belief updating—versus mentalize about them *more*—therefore generating alternative hypotheses to accommodate the surprising behavior.

To summarize, our proposal is that instances of social and moral decision-making that appear to be motivated may instead be compatible with Bayesian-rational reasoning. Our strong prior beliefs—e.g., positive beliefs about the ingroup—are often protected from revision through the generation of auxiliary hypotheses (Gershman, 2019). Further work is needed to differentiate between procedurally rational updating that appears irrational, and motivated updating that is driven by social and affective considerations (e.g., attachment to ingroup members). Finally, we call upon future work to examine the proximate and ultimate costs and benefits of motivated updating, above and beyond those of Bayesian updating.

Conclusion

We reviewed evidence that supports the engagement of mentalizing for optimal and suboptimal decision-making, in the contexts of moral judgment and economic exchange. People engage in accurate mentalizing, leading to social and non-social rewards, i.e., beating the competition, learning from others' strategies, and identifying cooperative partners. But, people can also engage in "biased" mentalizing, with the aim of protecting their positive impressions of close others (friends, ingroup members), leading to direct and indirect losses. Even so, as we discussed in the final section, seemingly "biased" decisions, i.e., discounting negative feedback about close others, may in fact be Bayesian-rational, stemming from differences in people's prior beliefs and knowledge. Determining which kinds of decisions reflect rational updating versus motivated reasoning will be an important question to address going forward. We look forward to future research, which will continue to enhance our understanding of how mentalizing contributes to value-based decision-making.

Acknowledgements We thank Josh Hirschfeld-Kroen, Emma Alai, and Simon Karg for their thoughtful feedback, and the John Templeton Foundation for support.

References

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*, 268–277.

Baron, S. G., Gobbini, M. I., Engell, A. D., & Todorov, A. (2011). Amygdala and dorsomedial prefrontal cortex responses to appearance-based and behavior-based person impressions. *Social Cognitive and Affective Neuroscience, 6*(5), 572–581.

Baumgartner, T., Götte, L., Gügler, R., & Fehr, E. (2012). The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. *Human Brain Mapping, 33*, 1452–1469.

Baumgartner, T., Schiller, B., Rieskamp, J., Gianotti, L. R. R., & Knoch, D. (2014). Diminishing parochialism in intergroup conflict by disrupting the right temporo-parietal junction. *Social Cognitive and Affective Neuroscience, 9*(5), 653–660.

Behrens, T. K. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature, 456*, 245–249.

Bhatt, M. A., Lohrenz, T., Camerer, C. F., & Montague, R. (2010). Neural signatures of strategic types in a two-person bargaining game. *Proceedings of the National Academy of Sciences of United States of America, 107*(46), 19720–19725.

Boorman, E. D., O'Doherty, J. P., Adolphs, R., & Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron, 80*, 1558–1571.

Brambilla, M., Carraro, L., Castelli, L., & Sacchi, S. (2019). Changing impressions: Moral character dominates impression updating. *Journal of Experimental Social Psychology, 82*, 64–73.

Bruguier, A. J., Quartz, S. R., & Bossaerts, P. (2010). Exploring the nature of "trader intuition". *The Journal of Finance, 65*(5), 1703–1723.

Cao, J., Kleiman-Weiner, M., & Banaji, M. R. (2019). People make the same Bayesian judgment they criticize in others. *Psychological Science, 30*(1), 20–31.

Coricelli, G., & Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences of United States of America, 106*(23), 9163–9168.

Cushman, F. A., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition, 127*(1), 6–21.

Cushman, F. (2020). Rationalization is rational. *Behavioral and Brain Sciences, 43*.

De Martino, B., O'Doherty, J. P., Ray, D., Bossaerts, P., & Camerer, C. (2013). In the mind of the market: Theory of mind biases value computation during financial bubbles. *Neuron, 79*, 1222–1231.

Decety, J., & Cacioppo, S. (2012). The speed of morality: A high-density electrical neuroimaging study. *Journal of Neurophysiology, 108*, 3068–3072.

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience, 8*(11), 1611–1618.

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., … Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Computational Biology, 10*, e1003810. https://doi.org/10.1371/journal.pcbi.1003810

Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Motivated moral reasoning. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. Burlington (Eds.), *The psychology of learning and motivation*. Burlington, Canada: Academic Press.

Evans, S., Fleming, S., Dolan, R. J., & Averbeck, B. B. (2011). Effects of emotional preferences on value-based decision-making are mediated by mentalizing and not reward networks. *Journal of Cognitive Neuroscience, 23*(9), 2197–2210.

Fareri, D. S., Chang, L. J., & Delgado, M. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience, 6*, 148.

Fareri, D. S., Chang, L. J., & Delgado, M. (2015). Computational substrates of social value in interpersonal collaboration. *The Journal of Neuroscience, 35*(21), 8170–8180.

Fischle, M. (2000). Mass response to the Lewinsky scandal: Motivated reasoning or Bayesian updating? *Political Psychology, 21*(1), 135–159.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50*, 531–534.

Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin and Review, 26*, 13. https://doi.org/10.3758/s13423-018-1488-8

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science, 24*(1), 38–44.

Güroğlu, B., van den Bos, W., Rombouts, S. A. R. B., & Crone, E. A. (2010). Unfair? It depends: Neural correlates of fairness in social context. *Social Cognitive and Affective Neuroscience, 5*(4), 414–423.

Gweon, H., Dodell-Feder, D., Bedny, M., & Saxe, R. (2012). Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. *Child Development., 83*, 1853–1868.

Hackel, L. M., & Amodio, D. M. (2018). Computational neuroscience approaches to social cognition. *Current Opinion in Psychology, 24*, 92–97.

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience, 18*(9), 1233–1235.

Hahn, U., & Harris, A. J. L. (2014). What does it mean to be biased: Motivated reasoning and rationality. In B. H. Ross (Ed.), *Psychology of learning and motivation*. San Diego, CA: Academic Press.

Hájek, A. (2008). Dutch book arguments. In P. Anand, P. Pattanaik, & C. Puppe (Eds.), *The handbook of rational and social choice*. New York, NY: Oxford University Press.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences of United States of America, 105*(18), 6741–6746.

Hare, T., Camerer, C. F., Knoepfle, D. T., O'Doherty, J. P., & Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *The Journal of Neuroscience, 30*(2), 583–590.

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes, 70*(2), 117–133.

Hill, C. A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J. P., & Ruff, C. C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience, 20*, 1142–1149.

Hill, S. J. (2017). Learning together slowly: Bayesian learning about political facts. *The Journal of Politics, 79*(4), 1403–1418.

Hughes, B. L., Zaki, J., & Ambady, N. (2017). Motivation alters impression formation and related neural systems. *Social Cognitive and Affective Neuroscience, 12*(1), 49–60.

Inbar, Y., Pizarro, D. A., & Cushman, F. (2012). Benefiting from misfortune: When harmless actions are judged to be morally blameworthy. *Personality and Social Psychology Bulletin, 38*(1), 52–62.

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron, 58*, 284–294.

Killen, M., Mulvey, K. L., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition, 119*, 197–215.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science, 308*(5718), 78–83.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia, 46*, 2949–2957.

Klucharev, V., Smidts, A., & Fernández, G. (2008). Brain mechanisms of persuasion: How 'expert power' modulates memory and attitudes. *Social Cognitive and Affective Neuroscience, 3*(4), 353–366.

Knobe, J. (2003). Intentional action and side-effects in ordinary language. *Analysis, 63*, 190–193.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron, 79*, 836–848.

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences of United States of America, 110*(14), 5648–5653.

Krueger, F., Grafman, J., & McCabe, K. (2008). Neural correlates of economic game playing. *Philosophical Transactions of The Royal Society, 363*, 3859–3874.

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., … Grafman, J. (2007). Neural correlates of trust. *Proceedings of the National Academy of Sciences of United States of America, 104*(50), 20084–20089.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In S. G. Harding (Ed.), *Can theories be refuted?* (pp. 205–259). Dordrecht, Netherlands: Springer.

Lee, D. (2006). Neural basis of quasi-rational decision making. *Current Opinion in Neurobiology, 16*(2), 191–198.

Leong, Y. C., & Zaki, J. (2017). Unrealistic optimism in advice taking: A computational account. *Journal of Experimental Psychology: General, 147*(2), 170.

Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Wheelright, S. J., Sadek, S. A., Suckling, J., … Baron-Cohen, S. (2009). Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience, 22*(7), 1623–1635.

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *The Journal of Neuroscience, 33*(50), 19406–19415.

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience, 8*(6), 623–631.

Mende-Siedlecki, P., & Todorov, A. (2016). Neural dissociations between meaningful and mere inconsistency in impression updating. *Social Cognitive and Affective Neuroscience, 11*(9), 1489–1500.

Mengotii, P., Dombert, P. L., Fink, G. R., & Vossel, S. (2017). Disruption of the right temporoparietal junction impairs probabilistic belief updating. *The Journal of Neuroscience, 37*(22), 5419–5428.

Molenberghs, P., Johnson, H., Henry, J. D., & Mattingley, J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience and Biobehavioral Reviews, 65*, 276–291.

Monroe, A. E., & Malle, B. F. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology, 116*(2), 215–236.

Park, B., Fareri, D. S., Delgado, M. R., & Young, L. (2019). How theory-of-mind brain regions process prediction error across relationship contexts (manuscript in preparation).

Park, B., & Young, L. (2020). An association between biased impression updating and relationship facilitation: A behavioral and fMRI investigation. *Journal of Experimental Social Psychology, 87*, 103916.

Phan, K. L., Sripada, C. S., Angstadt, M., & McCabe, K. (2010). Reputation for reciprocity engages the brain reward center. *Proceedings of the National Academy of Sciences of United States of America, 107*(29), 13099–13104.

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and-praise: The role of perceived metadesires. *Psychological Science, 14*(3), 267–272.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology, 39*, 653–660.

Park, B., Fareri, D., Delgado, M., & Young, L. (2020). The role of right temporoparietal junction in processing social prediction error across relationship contexts. *Social Cognitive and Affective Neuroscience, nsaa072*. https://doi.org/10.1093/scan/nsaa072

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications, 9*, 1027. https://doi.org/10.1038/s41467-018-03399-2

Rilling, J. K., Dagenais, J. E., Goldsmith, D. R., Glenn, A. L., & Pagnoni, G. (2008). Social cognitive neural networks during in-group and out-group interactions. *NeuroImage, 41*, 1447–1461.

Saxe, R. (2009). Theory of mind (neural basis). In W. Banks (Ed.), *Encyclopedia of consciousness*. Cambridge, MA: MIT Press.

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology, 55*, 87–124.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *Neuroimaging, 19*, 1835–1842.

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science, 17*(8), 692–699.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia, 43*(10), 1391–1399.

Saxe, R., Xiao, D. K., Kovacs, G., Perrett, D. I., & Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia, 42*, 1435–1446.

Schilbach, L., Eickhoff, S. B., Schultze, T., Mojzisch, A., & Vogeley, K. (2013). To you I am listening: Perceived competence of advisors influences judgment and decision-making via recruitment of the amygdala. *Social Neuroscience, 8*(3), 189–202.

Sellaro, R., Güroğlu, B., Nitsche, M. A., van den Wildenberg, W. P. M., Massaro, V., Durieux, J., … Colzato, L. S. (2015). Increasing the role of belief information in moral judgments by stimulating the right temporoparietal junction. *Neuropsychologia, 77*, 400–408.

Singer, T., & Fehr, E. (2005). The neuroeconomics of mind reading and empathy. *American Economic Review, 95*(2), 340–345.

Strang, S., Utikal, V., Fischbacher, U., Weber, B., & Falk, A. (2014). Neural correlates of receiving an apology and active forgiveness: An fMRI study. *PLoS One, 9*, e87654. https://doi.org/10.1371/journal.pone.0087654

Tannenbaum, D., Ditto, P. H. & Pizarro, D. A. (2008). *Different moral values produce different judgments of intentional action*. Poster Presented at the Annual Meeting of the Society for Personality and Social Psychology, Albuquerque, NM.

Thornton, M. A., & Mitchell, J. P. (2018). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex, 28*, 3505–3520.

Treadway, M. T., Buckholtz, J. W., Martin, J. W., Jan, K., Asplund, C. L., Ginther, M. R., … Marois, R. (2014). Corticolimbic gating of emotion-driven punishment. *Nature Neuroscience, 17*(9), 1270–1275.

Van Swol, L. M. (2009). The effects of confidence and advisor motives on advice utilization. *Communication Research, 36*(6), 857–873.

Walter, H., Adenzato, M., Ciaramidaro, A., Enrici, I., Pia, L., & Bara, B. G. (2004). Understanding intentions in social interaction: The role of the anterior paracingulate cortex. *Journal of Cognitive Neuroscience, 16*(10), 1854–1863.

Will, G.-J., Crone, E. A., & Güroğlu, B. (2015). Acting on social exclusion: Neural correlates of punishment and forgiveness of excluders. *Social Cognitive and Affective Neuroscience, 10*(2), 209–218.

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition, 100*, 283–301.

Yamada, M., Camerer, C. F., Fujie, S., Kato, M., Matsuda, T., Takano, H., … Takahashi, H. (2012). Neural circuits in the brain that are activated when mitigating criminal sentences. *Nature Communications, 3*, 759. https://doi.org/10.1038/ncomms1757

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 83*(2), 260–281.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of United States of America, 107*(15), 6753–6758.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of United States of America, 104*(20), 8235–8240.

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology, 1*, 333–349.

Young, L., & Saxe, R. (2009). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience, 21*(7), 1396–1405.

Young, L., Scholz, J., & Saxe, R. (2011). Neural evidence for "intuitive prosecution": The use of mental state information for negative moral verdicts. *Social Neuroscience, 6*(3), 302–315.

The Role of Morality in Social Cognition

Jennifer L. Ray, Peter Mende-Siedlecki, Ana Gantman, and Jay J. Van Bavel

Over the past few decades, two-factor models of social cognition have emerged as the dominant framework for understanding impression formation. Despite the differences in the labeling of the two factors, there is wide agreement that the core of one dimension reflects social/relational potential (which we will call sociability), and the other, competence/capacity (which we will call competence). However, scholars dating back to Aristotle have argued that morality may be the most important basis on which to form social evaluations, because competence and sociability could only be virtuous, sincere, and trustworthy if expressed through a moral character (MacIntyre, 1984). Indeed, recent work demonstrates that morality judgments influence the evaluation of other characteristics, shaping evaluations of the two core dimensions that dominate the literature in person perception: competence and sociability. In the current chapter, we will: (1) briefly describe several two-factor models of social cognition, (2) review evidence for why morality should be treated as a unique and primary dimension, and (3) discuss the flexibility of impression formation due to goals that affect attention to, and appraisal of, the moral dimension. This review reflects a growing consensus that two-factor models fail to capture the rich nature of impression formation (Tamir, Thornton, Contreras, & Mitchell, 2016; Thornton & Mitchell, 2018) and that morality might be the most important dimension of person perception (Goodwin, 2015).

J. L. Ray (✉)
Department of Psychology, New York University, New York, NY, USA
e-mail: Jennifer.Ray@nyu.edu

P. Mende-Siedlecki
University of Delaware, Newark, DE, USA

A. Gantman
City University of New York, New York, NY, USA

J. J. Van Bavel
New York University, New York, NY, USA

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_28

Models of Social Cognition

Although mentalizing is often described as thinking about transient mental states, there is a close, bi-directional relationship between inferring mental states and attributing stable traits**.** Indeed, some have argued that personality ought to be reconceptualized in terms of repeated action patterns and tendencies embedded in situations, which often depend on mental states such as expectations and beliefs regarding self-efficacy (Mischel, 1973). In this view, a trait we might consider, such as “ambitious,” is determined by our willingness to pursue our goals, which is in turn affected by self-efficacy beliefs activated in the relevant moment for that goal-directed behavior. People automatically infer traits from the behavior and mental states of others, a phenomenon called spontaneous trait inference (Todorov & Uleman, 2002, 2003; Winter & Uleman, 1984). The relationship between trait inferences and mental states is born out on a neural level as well. For example, recent work demonstrates that our neural representations of others reflect the mental states we believe they habitually experience (Thornton, Weaverdyck, & Tamir, 2019). Finally, we use our schemas of individuals to infer their mental states in a given situation (Higgins, Rholes, & Jones, 1977). Thus, when we mentalize, we make automatic judgments about others’ transient mental states and enduring traits; these judgments interact and inform one another, and so to understand the perception of short-lived mental states, we must also investigate the perception of stable traits.

There is a rich literature full of two-factor models to explain this essential element of social cognition. Indeed, contemporary psychology has seen the proliferation of two-dimensional models of self, interpersonal, and intergroup perception in a variety of literatures including social roles (Eagly & Steffen, 1984), stereotyping (Fiske, Cuddy, Glick, & Xu, 2002; Phalet & Poppe, 1997), dehumanization (Haslam, 2006), and mind perception (Gray, Gray, & Wegner, 2007). These models provide the foundation for the three-dimensional and dynamic approach we describe in this chapter. To provide the context for how and why a three-factor, dynamic model expands beyond previous work, we briefly review various two-factor models.

The two dimensions of these models are referred to by different names in each literature, including self-profitable-other profitable traits (Peeters, 1992), agency-communality (Eagly & Steffen, 1984); agency-communion (Abele & Wojciszke, 2007), or competence-warmth (Fiske et al., 2002). The first question calls attention to, and describes appraisal of, the core dimension *warmth*; the second question, to the core dimension *competence*. In this model, evaluations of morality and sociability are characterized as sub-components of the overarching warmth dimension. In a related conceptualization, Leach (2006) labels a first dimension *power* that encapsulates judgments of competence, strength, prestige, and activity, whereas the second dimension labeled *benevolence* encapsulates judgments of sociality, morality, cooperation, and compatibility. Importantly, both of these approaches treat traits associated with sociability and morality as pro-social and fold them into a single dimension (i.e., *warmth* or *benevolence*) (Leach et al., 2007; see also Fiske et al., 2002; Fiske, 2018).[1]

The literature on dehumanization and mind perception also relies on two-factor models. However, models in this literature differ in fundamental ways from the two-factor models previously described. In modeling humanness, Haslam (2006) proposes two distinct sets of attributes: (1) characteristics that are *uniquely human,* and (2) characteristics that are aspects of *human nature*. Uniquely human characteristics include refinement, civility, rationality, cultivation, and moral sensibility. In contrast, human nature includes emotional responsiveness, warmth, cognitive openness, agency, and depth. In this model, morality is found on the first dimension, whereas agency and warmth are found together on the second. The denial of each type of humanness results in a corresponding form of dehumanization—denying human uniqueness results in an implicit vertical comparison that the individual is sub-human, whereas denying human nature results in a horizontal separation that the individual is non-human (Haslam, 2006). Despite locating moral sensibility on a separate dimension from both agency and warmth, Haslam (2012) suggests that the moral status of diverse targets co-varies with the dimensions of humanness (Bastian, Laham, Wilson, Haslam, & Koval, 2011). In this way, the morality of targets appears to be conceived as an emergent property of the two dimensions of the model as opposed to being contained within one of them.

The two-factor model that focuses most explicitly on morality, however, argues that the two dimensions of mind perception—*agency* and *experience*—form the essence of moral judgment (Gray, Young, & Waytz, 2012). In this model, agency, or the perceived capacity to intend and act is orthogonal to experience, or the perceived capacity for sensations and feelings (Gray et al., 2007). In specifying the structure of mind perception, the authors tested 13 characters (types of living humans, non-human animals, a dead woman, God, and a sociable robot) against 18 capacities (Gray et al., 2007). Factor analysis revealed two dimensions: the *agency* dimension that included the capacities for planning, communication, thought, self-control, emotion recognition, memory, *and* morality; and the experience dimension that included the capacities for hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, and joy (Gray et al., 2007). Unlike most of the models previously described, the two-factor model of mind perception locates the capacity for morality as a sub-component of agency.

The relationships among agency, experience, and morality are complex because the authors also make links between perceiving minds along the agency and experience dimensions, and dyadic morality (Gray et al., 2012), or the attribution of moral

[1] Still other models emphasize a *socio-relational* dimension and a *competence* or achievement dimension (Wojciszke, 2005; Ybarra, Chan, & Park, 2001), or an *intended goal* dimension and *goal attainment* dimension (Phalet & Poppe, 1997). In the domain of face perception, trustworthiness and dominance are seen as the two core dimensions (see Todorov, 2008), though note that (a) other work using stimuli that vary more substantially on age obtains a third dimension, related to youthfulness/attractiveness (Vernon, Sutherland, Young, & Hartley, 2014), and further, (b) more recent work suggests that rather than reflecting fixed trait dimensions, social face evaluation is best captured by a dynamic and contextually sensitive framework (Stolier, Hehman, & Freeman, 2018).

rights and moral responsibilities to those minds. Specifically, perceived agency qualifies an entity as capable of doing good or evil, whereas perceived experience qualifies an entity as benefiting from good or suffering from evil (Gray et al., 2012). According to this model, moral judgment depends on a dyadic template of two minds: a moral agent and a moral patient (i.e., the action of the moral agent and the resultant suffering or salvation of the moral patient) (Gray et al., 2012). In other words, perceptions of moral capacity and other competence-related capacities produce perceptions of agency, perceived agency triggers the ascription *moral agent*, and the ascription *moral agent* disqualifies the ascription *moral patient*. At once, targets' perceived capacity for morality informs perceptions of their agency, and simultaneously informs whether they're perceived as good versus bad, or whether their actions are perceived as right versus wrong.

In the next section, we review evidence suggesting that morality is not only a critical and separable dimension of social cognition, but that it may even be the primary dimension. Decades of research have identified the centrality of the warmth and competence dimensions (e.g., Rosenberg, Nelson, & Vivekananthan, 1968), and suggest the universality of these dimensions across cultures (Fiske, Cuddy, & Glick, 2007). Moreover, even in two-factor models, one dimension often takes primacy over the other. For instance, the functionalist perspective invokes evolutionary reasons for the primacy of warmth over competence in social perception. Survival in the social world requires that in encounters with others, individuals must immediately determine whether the target has beneficial or harmful intentions, and only later, whether the target has the ability to enact his or her intentions (Fiske et al., 2007). Along similar lines, goal-oriented approaches to person perception suggest that approach-avoidance underlies impression formation and is more directly based on appraisals of warmth than competence (Wojciszke & Abele, 2008; Wojciszke, Bazinska, & Jaworski, 1998). However, when morality is folded into sociability (or within agency), its independence as a third dimension of social perception, and perhaps the *primary* dimension of social perception, remains obscured.

Morality: The Third Dimension of Social Cognition

Are moral evaluations different from other types of evaluations? Is judging a target to be moral different than judging a target to be merely an effective, competent, or sociable one? Aristotle saw morality as the most important basis on which to form positive evaluations, because competence and sociability could only be virtuous, sincere, and trustworthy if expressed through a moral character (MacIntyre, 1984). Aristotle's perspective suggests that morality judgments can color the perception of every other characteristic—consider that immoral competence is dangerous and immoral sociability is disingenuous (Leach et al., 2007; Osgood, Suci, & Tannenbaum, 1957). Accordingly, it may be impossible to form *meaningful* evaluations of competence or sociability without first determining whether the target is a moral or immoral actor (see Goodwin, Piazza, & Rozin, 2014; Landy, Piazza, & Goodwin, 2016). In this section, we review findings relevant to the moral judgment of targets (and actions), demonstrating that moral judgments appear to be faster and more extreme, privileged, sticky, and affect-laden than non-moral ones, as well as able to shape non-moral ones. We present evidence that moral evaluations are different from other types of evaluation, and in ways that motivate our belief that they may serve a primary role in social cognition.

As suggested in the previous section, morality and sociability information have often been combined into one superordinate dimension, which is often referred to as warmth (e.g., Fiske et al., 2007). However, work over the past decade has begun to tease these dimensions apart, making a distinction between traits associated with morality (e.g., honesty, sincerity, and trustworthiness) (Brambilla, Rusconi, Sacchi, & Cherubini, 2011), as well as fairness, loyalty, courage, etc. (Goodwin et al., 2014), versus traits more generally reflective of sociability (e.g., likability, warmth, and friendliness) (Brambilla et al., 2011), as well as extroversion, agreeableness, enthusiasm, etc. (Goodwin et al., 2014).

Taken together, this work strongly suggests that morality can be dissociated from sociability. For example, a three-factor measurement model that included morality alongside competence and sociability accounted for participants' evaluations of groups better than more parsimonious alternatives (Leach et al., 2007). Specifically, morality was more important to individuals' positive evaluations of their in-group than either competence or sociability. Only morality affected participants' levels of pride in, or social distancing from, their in-group (Leach et al., 2007). Subsequent research has both replicated the primacy of morality in the context of group impression formation (Brambilla, Cherubini, & Yzerbyt, 2012), and extended this observation to information processing at the level of individual targets. Specifically, when forming global impressions of other individuals, participants focused more on traits related to morality (like "trustworthy"), versus traits related to either sociability (like "likable") or competence (like "intelligent," Brambilla et al., 2011). In addition, morality dominates over sociability and competence information when *updating* impressions as well (Brambilla, Carraro, Castelli, & Sacchi, 2019).

Further work has observed the primacy of morality in person perception, and its dis-sociability from warmth (e.g., sociability). For example, across seven studies, Goodwin et al. (2014) confirmed the domains of morality and warmth are perceived as being distinct from one another, and further, that the moral character associated with a given trait was a better predictor of judgments of its relevance to one's identity, its desirability, and its controllability as compared to warmth. Further, information related to morality was a better predictor of overall, global impressions of targets than information related to warmth (Goodwin et al., 2014).

Echoing the Aristotelian perspective noted at the beginning of this section, even more recent work indicates that the primacy of morality is such that positive evaluations of both sociability and competence hinge on the presence or absence of moral character (Landy et al., 2016). For example, sociable and competent targets were only rated positively if they were also moral,[2] whereas moral targets were always rated positively, independent of any other trait information. Taken together, this

[2] Interestingly, work from the same authors (Piazza, Goodwin, Rozin, & Royzman, 2014) suggests that morality itself may be further parcellated into core goodness traits that amplify moral goodness unconditionally versus value commitment traits (e.g., like "dedicated") that can amplify moral goodness *or* moral badness conditionally.

work provides the basis for the morality dominance hypothesis. While other work has demonstrated that preferences for moral versus immoral traits may be dynamic and sensitive to context (Melnikoff & Bailey, 2018), morality information drives impression formation, and moral traits are viewed positively—at least to the extent that a perceiver views themselves as being moral (Landy, Piazza, & Goodwin, 2018).

Morality is also separate from and primary to competence and sociability (Brambilla et al., 2011, 2019; Brambilla et al., 2012; Goodwin et al., 2014; Landy et al., 2016). Work testing the dominance of moral appraisals over competence appraisals demonstrated the chronic accessibility of morality traits as compared to competence traits, as well as that ascriptions of moral traits better predicted global impressions of targets than the ascription of competence traits (Wojciszke et al., 1998). In a related study, the authors tested how both positive and negative morality and competence information was integrated to form impressions of targets. The authors predicted that morality information would have a greater than additive effect on global impressions of targets. The results showed that the evaluative meaning of moral information was independent of competence information (i.e., the moral information retained its direct effect on impression formation irrespective of the target's competence), whereas the opposite was found for competence information. The effect of competence information on global impressions was shaped by the positivity versus negativity of moral information. In other words, morality information provided the necessary context for evaluating whether an individual's competence positively or negatively impacts a global impression of the target, but the same was not true for competence information (Wojciszke et al., 1998), which suggests that moral information is primary.

Research on the how individuals recall information about targets, as well as how individuals revise their prior impressions of targets when they receive more information, also underscores the primacy of morality. For example, researchers explored how participants would form an impression when the behavioral information (related to both morality and competence) at Time 2 contradicted the behavioral information they received at Time 1 (also related to both morality and competence, but opposite in valence from Time 1; Ybarra, 2001). The results revealed that the amount of change in impressions of the target was greater for the moral domain than the competence domain (when negative behavioral information followed positive behavioral information). These results highlight the greater influence of negative moral information over negative competence information on impression revision processes, when the initial evaluation is positive and more susceptible to adaptation (Ybarra, 2001). Moreover, participants recalled morality information better than either intelligence or neutral information (De Bruin & Van Lange, 1999). We recognize that alone these studies are consistent with a two-factor model (in which morality and sociability are not distinct). But evidence that morality judgments trump sociability judgments and competence judgments suggests that it is weighted most heavily in impression formation.

There are also multiple theoretical accounts suggesting differential processing of moral versus competence information (see Brambilla et al., 2011). One possibility is that perceivers have a lay theory that everyone can behave in moral ways, but only immoral individuals behave in immoral ways. In other words, moral behaviors are less informative about traits, because they may not only reveal underlying characteristics but also reflect situational constraints on behavior. Conversely, perceivers have a lay theory that only competent individuals are capable of great successes, whereas everyone can behave incompetently at one time or another due to situational constraints. Thus, incompetent behaviors are less informative about traits in this domain, whereas competent behaviors are more diagnostic (i.e., they tell you more about the person you are perceiving). These differential expectations about trait-behavior relations produce a negativity bias in the moral domain because immoral behaviors are more informative, and a positivity bias in the competence domain because competent behaviors are more informative.

A final reason to consider morality as a primary dimension in person perception is that moral evaluations are typically affect-laden (Haidt, 2001), and morality is salient in visual perception (Gantman & Van Bavel, 2015). Affectively valenced evaluations are made particularly rapidly (Zajonc, 1980) and moral intuitions are especially affect-laden (Haidt, Koller, & Dias, 1993), suggesting that morally laden rapid inferences of others may occur more quickly than inferences about sociability and competence (though see Fiske et al., 2007). Specifically, within the domain of impression formation, predictions of a target's future behavior in the moral domain were based in part on the perceiver's affective responses to the target, whereas predictions in the competence domain were based only on ascriptions of a relevant personality trait (Brycz & Wojciszke, 1992). In the domain of perception, morally relevant words are more likely to be seen than matched non-moral words (Gantman & Van Bavel, 2014). Together, this evidence is suggestive though not conclusive that morally relevant information may be processed prior to competence and sociability judgments. Empirical evidence demonstrating that moral evaluations occur earlier than other person perception judgments is a promising avenue for future research.

Discussion

We have sought to review evidence that morality should be added to two-factor models of person perception, creating a three-factor model. In this view, moral information exerts a powerful influence on social cognition and influences evaluations within the other dimensions, like sociability and competence. This approach helps account for a wide literature on impression formation and mental state attribution while raising exciting possibilities for future research. In the final section, we highlight gaps in the current literature and propose some potential avenues for future work in the domain of morality and impression formation.

Moral character judgments may be primary to other types of moral judgments, such as blame and praise, which fundamentally depend on the mental states of the agents. For instance, blame is especially sensitive to intent (Malle, Guglielmo, & Monroe, 2014). Some psychologists have begun to advance a "person-based" as opposed to "act-based" theory of moral blame. They have demonstrated that moral blame can be disproportionate to the actual harm caused by an agent and that relatively harmless acts can receive harsh moral judgments (Pizarro & Tannenbaum, 2011). From this perspective, psychological theories of moral judgment are incomplete because they disregard the primacy of moral character evaluations (e.g., underlying traits, dispositions, and character) to assigning blame and praise, and instead narrowly focus on the local features of the act and agent (e.g., whether the action violates a rule or whether the agent's mental state at the time of the action allowed for alternative actions, or whether the act caused harm) (Pizarro & Tannenbaum, 2011). For instance, ascriptions of blame to an actor for a car accident were significantly heightened when the actor's underlying motive for speeding was socially undesirable (to hide a vial of cocaine) versus socially desirable (to hide an anniversary present for his parents; Alicke, 1992). Likewise, rashness leading to an immoral decision was seen as reflecting the wrongdoer's immoral underlying character, which intensified blame (Critcher, Inbar, & Pizarro, 2013). More work should examine how evaluations of a target's morality might shape perceptions of the target's competence or agency.

Moral judgments of an action can also impact whether that action is perceived as having been done intentionally (also known as the Side Effect Effect; Knobe, 2003). In this work, participants learn about a Vice President of a company who approaches the Chairman of the Board with a project that will increase profits but, as a side effect, will either harm or help the environment. The Chairman of the Board says that he only cares about increasing profits, not the environment. People perceive the harm, but not the help to the environment, as intentional (Knobe, 2003). This underscores that moral character judgments affect perceptions of the agent's mental state. Yet little work has examined how these mental state attributions are allocated across multiple moral agents. For instance, how is moral blame allocated to the Vice President as compared to the Chairman of the Board? And how do other mental state inferences, about deliberation vs. implementation, play into these moral judgments? Given that many moral actions are distributed across multiple moral agents, this will be a fruitful area for future research.

As well, very little work has examined how these judgments unfold over time. Like other evaluations, impressions are likely formed through a series of cycles that occur over time with evaluations being updated and adjusted due to contextual and motivational information (Van Bavel, Xiao, & Cunningham, 2012). As such, the final judgment of the morality, competence, and sociability of a target is an emergent property of multiple processes unfolding over time (Cunningham & Zelazo, 2007; Cunningham, Zelazo, Packer, & Van Bavel, 2007; Ferguson & Wojnowicz, 2011; Freeman & Ambady, 2011; Scherer, 2009; Van Bavel et al., 2012). It is unknown how appraisals in each domain might influence appraisal in the others over time. For instance, do moral judgments emerge first in time? If so, this would provide yet another form of moral precedence and might constrain subsequent competence and sociability judgments.

This dynamic approach to social cognition also assumes that context, motivation, and goals will shape attention to, appraisal of, and affective reactions induced by behavioral information about targets, because information processing in general is highly flexible and dependent on both cognitive and motivational goals (Hilton & Darley, 1991; Wojciszke, 2005). Simply put, goals guide information selection. Thus, perceivers appear to be highly interested in a target's morality in the absence of a particular goal. But, when a domain-specific goal is made clear, the task domain instead appears to drive the perceiver's search for corresponding information about the target (Wojciszke, 2005). For instance, when goals are related to competence (e.g., in hiring decisions), perceivers' attention is more directed to competence information. However, when goals are focused on moral judgment (e.g., dating decisions), perceivers' attention is more focused on moral information (see Everett, Pizarro, & Crockett, 2016). Moreover, the value placed on moral vs. immoral traits likely depends on their motivational relevance (Melnikoff & Bailey, 2018). Beyond goals, individual differences are also associated with chronically accessible constructs that perceivers use when evaluating targets and likely play a role in selective impression formation (Wojciszke et al., 1998). Finally, identical actions can be construed in both moral and competence terms depending on the distinct features of the action to which the perceiver is attending (Wojciszke, 1994), and this has important consequences for subsequent evaluations (Van Bavel et al., 2012).

Conclusion

Social cognition is shaped by the *interplay* of three distinct dimensions of morality, sociability, and competence, but moral states and traits may be more relevant to inferring whether a target represents a threat or opportunity than either sociability or competence. Inferences about morality exert their primacy, either by overpowering or by shaping judgments of sociability and competence. It may also be the case that the primacy of morality affects mental state ascriptions. Fully understanding the nature of impression formation, person perception, and social cognition is severely impoverished without understanding the power of morality.

References

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology, 93*(5), 751–763.

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology, 63*, 368–378.

Bastian, B., Laham, S. M., Wilson, S., Haslam, N., & Koval, P. (2011). Blaming, praising, and protecting our humanity: The implications of everyday dehumanization for judgments of moral status. *British Journal of Social Psychology, 50*(3), 469–483.

Brambilla, M., Carraro, L., Castelli, L., & Sacchi, S. (2019). Changing impressions: Moral character dominates impression updating. *Journal of Experimental Social Psychology, 82*, 64–73.

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology, 41*(2), 135–143.

Brambilla, M., Sacchi, S., Rusconi, P., Cherubini, P., & Yzerbyt, V. Y. (2012). You want to give a good impression? Be honest! Moral traits dominate group impression formation. *British Journal of Social Psychology, 51*(1), 149–166.

Brycz, H., & Wojciszke, B. (1992). Personality impressions on ability and morality trait dimensions. *Polish Psychological Bulletin, 23*(3), 223–236.

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science, 4*(3), 308–315.

Cunningham, W. A., & Zelazo, P. D. (2007). Attitudes and evaluations: A social cognitive neuroscience perspective. *Trends in Cognitive Sciences, 11*(3), 97–104.

Cunningham, W. A., Zelazo, P. D., Packer, D. J., & Van Bavel, J. J. (2007). The iterative reprocessing model: A multilevel framework for attitudes and evaluation. *Social Cognition, 25*(5), 736–760.

De Bruin, E. N., & Van Lange, P. A. (1999). Impression formation and cooperative behavior. *European Journal of Social Psychology, 29*(2–3), 305–328.

Eagly, A. H., & Steffen, V. J. (1984). Gender stereotypes stem from the distribution of women and men into social roles. *Journal of Personality and Social Psychology, 46*, 735.

Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General, 145*, 772–787.

Ferguson, M. J., & Wojnowicz, M. T. (2011). The when and how of evaluative readiness: A social cognitive neuroscience perspective. *Social and Personality Psychology Compass, 5*, 1018–1038.

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science, 27*(2), 67–73.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*, 77–83.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878–902.

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review, 118*, 247–279.

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science, 24*(1), 38–44.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology, 106*(1), 148–168.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619–619.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101–124.

Gantman, A. P., & Van Bavel, J. J. (2014). The moral pop-out effect: *Enhanced perceptual awareness of morally relevant stimuli. Cognition, 132*(1), 22-29.

Gantman, A. P., & Van Bavel, J. J. (2015). Moral perception. *Trends in Cognitive Sciences, 19*(11), 631-633.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65*(4), 613–628.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10*, 252–264.

Haslam, N. (2012). Morality, mind, and humanness. *Psychological Inquiry, 23*(2), 172–174.

Higgins, E., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology, 13*, 141–154.

Hilton, J. L., & Darley, J. M. (1991). The effects of interaction goals on person perception. *In Advances in experimental social psychology* (Vol. 24, pp. 235-267). Academic Press.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*(3), 190–194.

Landy, J. F., Piazza, J., & Goodwin, G. P. (2016). When it's bad to be friendly and smart: The desirability of sociability and competence depends on morality. *Personality and Social Psychology Bulletin, 42*(9), 1272–1290.

Landy, J. F., Piazza, J., & Goodwin, G. P. (2018). Morality traits still dominate in forming impressions of others. *Proceedings of the National Academy of Sciences, 115*(25), E5636–E5636.

Leach, C. W. (2006). *The meaning of prejudice* (Unpublished manuscript). University of Sussex, Brighton, England.

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology, 93*(2), 234–249.

MacIntyre, A. C. (1984). *After virtue* (Vol. 211). Notre Dame, IN: University of Notre Dame Press.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*, 147–186.

Melnikoff, D. E., & Bailey, A. H. (2018). Preferences for moral vs. immoral traits in others are conditional. *Proceedings of the National Academy of Sciences, 115*, E592–E600.

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80*, 252–283.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Peeters, G. (1992). Evaluative meanings of adjectives in vitro and in context: Some theoretical implications and practical consequences of positive negative asymmetry and behavioral-adaptive concepts of evaluations. *Psychologica Belgica, 32*, 211–231.

Phalet, K., & Poppe, E. (1997). Competence and morality dimensions of national and ethnic stereotypes: A study in six eastern-European countries. *European Journal of Social Psychology, 27*, 703–723.

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 91–108). Washington, DC: American Psychological Association.

Piazza, J., Goodwin, G. P., Rozin, P., & Royzman, E. B. (2014). When a virtue is not a virtue: Conditional virtues in moral evaluation. *Social Cognition, 32*(6), 528-558.

Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multi-dimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology, 9*, 283–294.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion, 23*, 1307–1351.

Stolier, R. M., Hehman, E., & Freeman, J. B. (2018). A dynamic structure of social trait space. *Trends in Cognitive Sciences, 22*(3), 197–200.

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences, 113*(1), 194–199.

Thornton, M. A., & Mitchell, J. P. (2018). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex, 28*(10), 3505–3520.

Thornton, M. A., Weaverdyck, M. E., & Tamir, D. I. (2019). The brain represents people as the mental states they habitually experience. *Nature Communications, 10*, 2291.

Todorov, A. (2008). Evaluating faces on trustworthiness: An extension of systems for recognition of emotions signaling approach/ avoidance behaviors. In A. Kingstone & M. Miller (Eds.), *Annals of the New York Academy of Sciences. The year in cognitive neuroscience 2008* (Vol. 1124, pp. 208–224). Hoboken, NJ: Blackwell Publishing.

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: Evidence from a false recognition paradigm. *Journal of Personality and Social Psychology, 83*(5), 1051.

Todorov, A., & Uleman, J. S. (2003). The efficiency of binding spontaneous trait inferences to actors' faces. *Journal of Experimental Social Psychology, 39*(6), 549–562.

Van Bavel, J. J., Xiao, Y. J., & Cunningham, W. A. (2012). Evaluation is a dynamic process: Moving beyond dual system models. *Social and Personality Psychology Compass, 6*(6), 438–454.

Vernon, R. J., Sutherland, C. A., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences, 111*(32), E3353–E3361.

Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology, 47*(2), 237–252.

Wojciszke, B. (1994). Multiple meanings of behavior: Construing actions in terms of competence or morality. *Journal of Personality and Social Psychology, 67*(2), 222–232.

Wojciszke, B. (2005). Morality and competence in person-and self-perception. *European Review of Social Psychology, 16*(1), 155–188.

Wojciszke, B., & Abele, A. E. (2008). The primacy of communion over agency and its reversals in evaluations. *European Journal of Social Psychology, 38*(7), 1139–1147.

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin, 24*(12), 1251–1263.

Ybarra, O. (2001). When first impressions don't last: The role of isolation and adaptation processes in the revision of evaluative impressions. *Social Cognition, 19*(5), 491–520.

Ybarra, O., Chan, E., & Park, D. (2001). Young and old adults' concerns about morality and competence. *Motivation and Emotion, 25*(2), 85–100.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35*(2), 151–175.

An Interbrain Approach for Understanding Empathy: The Contribution of Empathy to Interpersonal Emotion Regulation

S. Franklin-Gillette and S. G. Shamay-Tsoory

Introduction

Empathy refers to our ability to understand or to share the experience of others. Most models of empathy divide empathy into two factors: emotional empathy, i.e., feeling the emotions of others, and cognitive empathy (also referred to as mentalizing), i.e., understanding the thoughts and motivations of others (Cuff, Brown, Taylor, & Howat, 2016; Gonzalez-Liencres, Shamay-Tsoory, & Brüne, 2013; Smith, 2006). Emotional empathy is believed to be a more spontaneous, lower-order phenomenon (evolutionarily wise) than cognitive empathy (de Waal & Preston, 2017; Shamay-Tsoory, 2011), while cognitive empathy requires higher-order cognitive abilities, such as theory of mind (de Waal & Preston, 2017; Smith, 2006). Correspondingly, research shows that these subtypes of empathy operate independently, and differ from one another at the neurological, as well as the behavioral, levels (Cuff et al., 2016; Eres, Decety, Louis, & Molenberghs, 2015; Shamay-Tsoory, 2011; Shamay-Tsoory, Aharon-Peretz, & Perry, 2009; Singer & Klimecki, 2014; Zaki, Bolger, & Ochsner, 2008).

One significant role of both emotional and cognitive empathy is in interpersonal regulation of distress, a prominent feature of social life, as individuals frequently turn to friends and family when overwhelmed by negative emotions (Zaki & Williams, 2013). Although the literature on empathy is based largely on the premise that it plays a major role in alleviating others' distress (Lamm & Silani, 2013), little research directly explored the contribution of empathy to the distress regulation of others, especially with regard to the differentiating roles of the two different types

S. Franklin-Gillette (✉) · S. G. Shamay-Tsoory
Department of Psychology, University of Haifa, Haifa, Israel

The Integrated Brain and Behavior Research Center (IBBR), University of Haifa, Haifa, Israel
e-mail: mmalul1@univ.haifa.al.il

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_29

of empathy (emotional and cognitive). Indeed, despite its inherently interpersonal nature, research on empathy seems to remain primarily focused on the internal mechanisms of empathy (Main, Walle, Kho, & Halpern, 2017; Zaki et al., 2008), without exploring its outcome or function (Main et al., 2017). Given that empathy is identified as evolutionarily beneficial, and that the purpose of empathy is to alleviate the distress of a suffering target, this shift in research is crucial.

Emotion Regulation: Self-Regulation and Interpersonal Regulation

Emotion regulation refers to the ways individuals modulate their experience and expression of emotion, and is a widely studied construct in psychology. It includes how individuals influence the duration, intensity, and expression of their positive and negative emotions, both consciously and unconsciously (Beauregard, Lévesque, & Paquette, 2004; Gross & John, 2003). Effective emotion regulation is crucial for individuals to cope with stress, achieve their goals, interact socially, and adapt to their environments. Failure to appropriately regulate emotions is implicated in numerous psychiatric disorders, such as major depression and borderline personality disorder (Beauregard et al., 2004; Shipman, Schneider, & Brown, 2004). There are different discrete strategies of emotion regulation, many of which are attention-based strategies which rely on types of distraction. Simpler strategies in this group include avoiding distressing stimuli, e.g., through closing one's eyes or walking away, while more cognitively advanced strategies include distracting oneself by thinking about something else. Another common strategy is reappraisal—changing the way one thinks about a situation, thus changing its emotional impact (Gross & John, 2003). Suppression is another strategy, which involves inhibiting the emotion (Gross & John, 2003). These strategies are used to varying degrees of effectiveness in different situations (Gross & John, 2003; Webb, Miles, & Sheeran, 2012).

Neural models of emotion regulation frequently include the prefrontal cortex (PFC) and limbic system (Beauregard et al., 2004). The limbic system (including brain regions such as the amygdala) is implicated in emotional arousal, whereas different discrete brain regions are associated with varying regulation strategies. For example, reappraisal involves activations in the frontoparietal executive network and in cognitive control regions that dampen the amygdala, while unconscious regulation of fear involves the ventral anterior cingulate cortex (ACC) and the ventromedial PFC (Buhle et al., 2014; Etkin, Büchel, & Gross, 2015).

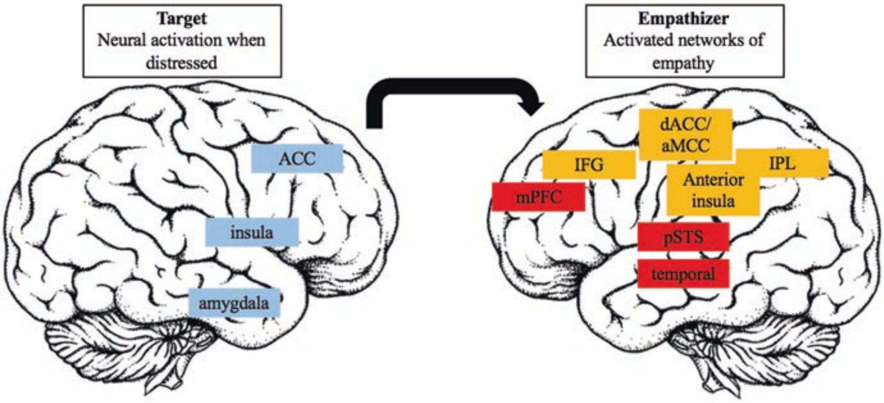
While social experiences in childhood and adulthood (Zaki & Williams, 2013) carry a great influence over how individuals regulate their emotions, most research focuses on how individuals manage their own emotions (intrapersonal emotion regulation). However, interpersonal emotion regulation is an important construct to study for many reasons. Research suggests that parents regulating their children's emotions helps children learn self-regulation (Reeck, Ames, & Ochsner, 2016).

Further, interpersonal emotion regulation is important for individuals who have difficulties regulating their own emotions, due to different psychopathologies (Reeck et al., 2016). Other research theorizes that interpersonal emotion regulation could be more effective than intrapersonal emotion regulation, as outside observers may have more objective views of a problem than a distressed individual, and therefore might be able to select a more effective regulation strategy, as they are not distracted by their emotions (Horn & Maercker, 2016; Levy-Gigi & Shamay-Tsoory, 2017).

Accordingly, a growing body of work now examines interpersonal emotion regulation—how individuals influence and affect the emotions of another (Reeck et al., 2016). In models of interpersonal emotion regulation (e.g., Niven, 2017; Reeck et al., 2016; Zaki & Williams, 2013), there is both a target—the individual experiencing emotion, and an observer/regulator, who affects the emotions of the target. This is important, as this process is inherently interpersonal, and must fully address the roles of both individuals involved. While the presence of others can affect an individual's emotions regardless of what those others do or do not do, many models require that the observer deliberately attempt to change the emotions of the target in order for the interaction to be considered interpersonal emotion regulation (Reeck et al., 2016). The target experiences a (usually distressing) emotion, which evokes a response in the observer. The observer then decides to and attempts to modulate the emotions of the target and can use a variety of strategies to do so. This creates the cyclical nature of interpersonal emotion regulation, noted in most models (Reeck et al., 2016; Zaki & Williams, 2013).

Empathy and Interpersonal Emotion Regulation

Zaki and Williams (2013) posit that empathy is evident in the interpersonal emotion regulation cycle, as the target evokes an empathetic reaction in the observer/regulator. Empathy may also serve as the observer's motive in changing the target's emotions. Similarly, Reeck et al.'s (2016) model of social regulation implicates brain regions that are associated with empathy: the dorsal premotor regions, dorsal medial PFC and precuneus in the regulator, and the TPJ, dorsal medial orbitofrontal cortex (OFC) and precuneus in the target. These regions are assumed to be involved in multiple stages of interpersonal emotion regulation: the regulator identifying the target's emotions, determining that regulation is necessary, and choosing how to regulate the target's emotions. Yet, the few models linking empathy with interpersonal emotion regulation seem to neglect the multi-faceted complexity of empathy, not addressing the fact that both emotional and cognitive empathy may contribute to the reduction of distress or pain in the target through different mechanisms. Here we propose a model that accounts for the neurological and behavioral differences of emotional and cognitive empathy, and the implications of those differences in the interpersonal emotion regulation process.



In order for the empathizer to feel empathy, he or she must know that the target is distressed. The target can either verbally express distress or convey negative emotions through body language and facial expressions. These negative emotions result from a neural network of distress based in the amygdala, insula, and ACC (i.e., Colibazzi et al., 2010). There are multiple paths through which the target's distress can trigger empathy, all of which are discussed in the following paragraphs.

Emotional Empathy and Interpersonal Emotion Regulation

Emotional empathy is an unconscious, automatic phenomenon that involves affective sharing—feeling the physical or emotional distress of another. Although empathy is not limited to distress, empathy in relation to interpersonal emotion regulation is specific to sharing and regulating distress. Its three main components are emotional contagion (feeling the same emotion as the target), emotion recognition (identifying what emotion the target is experiencing), and shared pain (feeling the same physical pain as the target) (Shamay-Tsoory, 2011). It can further be divided into the subcategories of personal distress (also called empathetic distress)—self oriented and focused on alleviating one's own pain, and empathetic concern (also called compassion)—feeling sympathy for another person (Eisenberg & Eggum, 2009; Singer & Klimecki, 2014). The primary difference between these two subcategories is that an individual experiencing personal distress is focused on their own distress, caused by the target, whereas an individual experiencing empathetic concern is focused on the distress of the target. While empathy is largely considered adaptive, there is evidence that empathetic personal distress is maladaptive, as an individual can become too overwhelmed with their own distress to engage socially, leading to withdrawal (Grynberg & Lopez-Perez, 2018). Empathetic concern, unlike personal distress, often leads to prosocial behavior and is a crucial aspect of social interaction, as it allows individuals to understand others and interact with them in positive, supportive ways (Berhardt & Singer, 2012).

Previous neuroimaging studies suggest that emotional empathy is based in shared neural networks. Observing others' emotions activates the same brain regions activated by experiencing emotions and pain (Berhardt & Singer, 2012). This is likely due to the mirror neuron system—a neural mechanism activated by observing others and neurologically mirroring their neural activation (Rizzolatti & Craighero, 2004). This system conceptually takes place, on a conceptual level, across different regions of the brain. The perception-action (or observation-execution) hypothesis states that perceiving an action likely activates neural representations of that action in the observer (de Waal & Preston, 2017; Shamay-Tsoory, 2011). Given that neural activations for emotions and pain are mirrored, numerous neural structures are implicated in emotional empathy depending on the specific emotion observed. However, some specific brain regions seem to be frequently activated. Research on the mirror neuron system has placed the center of this system in the inferior frontal gyrus (IFG) and the inferior parietal lobule (IPL) (Budell, Jackson, & Rainville, 2010; Shamay-Tsoory, 2011), thus implicating these regions in emotional contagion and mimicry. As emotional contagion is a basic part of emotional empathy, these regions play an important role in emotional empathy. Other regions, primarily the anterior insula (AI) and the dorsal-anterior/anterior-midcingulate cortex (dACC/aMCC) are activated in both experienced and observed pain, thus playing a role in emotional empathy for physical pain (Berhardt & Singer, 2012; Shamay-Tsoory, 2011). The motor cortex was also implicated in empathy in different EEG studies investigating the role of the suppression of mu rhythms, which occur only in the motor cortex. Suppression of mu rhythms appears in response to observing others in physical pain, as well as in recognition of positive and negative facial expressions (Moore, Gorodnitsky, & Pineda, 2012; Yang, Decety, Lee, Chen, & Cheng, 2009). This further demonstrates the neural base of empathy for pain and emotion, as neural activation in an empathizer mirrors the neural activation of a target.

One of the behavioral outcomes of the mirror neuron system is physical mimicry of facial expressions, body language, and verbal expression of emotions. Higher levels of empathy, frequently measured by questionnaire, are linked with embodied experiencing of observed emotions, aided and accompanied by automatic physical mimicry (Jospe, Flöel, & Lavidor, 2018). Accordingly, when observing pictures of people with happy or angry expressions, individuals higher in questionnaire-assessed emotional empathy tend to have more animated facial expressions than individuals lower in emotional empathy (Dimberg, Andréasson, & Thunberg, 2011). This physical mimicry of emotion has clear implications for interpersonal emotion regulation—seeing one’s own emotions mirrored in another could relieve distress through the mechanism of feeling understood, which activates the ventral striatum and middle insula—brain regions associated with reward. Additionally, synchrony in any emotion, positive or negative, activates the medial OFC and ventromedial PFC, brain regions associated with reward processing, as well as increasing feelings of closeness (Kuhn et al., 2010; Kühn et al., 2011).

There is applied research demonstrating the psychological benefits of emotional synchrony. For example, both psychological and behavioral synchrony between mothers and their infants predicts higher levels of self-regulation in early childhood and higher levels of empathy in late childhood and early adolescence, as well as reduction in psychosocial problems (Feldman, 2007a, 2007b). This synchrony is a form of co-regulation, as the infant learns from the mother how to modulate affective states in relation to social interactions (Feldman, 2007a, 2007b). Additionally, nonverbal motor synchrony between therapists and their clients is linked with higher quality therapeutic relationships and greater reductions of psychological symptoms. This nonverbal synchrony was suggested to be indicative of embodiment of the other's emotional state, which was predictive of treatment outcome (Ramseyer & Tschacher, 2011).

Another behavior that frequently results from empathy is social touch, involving physical consoling and comforting touch between individuals (De Waal, 2008). This is an integral aspect of social communication (Goldstein, Shamay-Tsoory, Yellinek, & Weissman-Fogel, 2016) and a behavioral manifestation of sympathy (Hertenstein & Weiss, 2011). It was found that individuals are capable of distinguishing the discrete emotions (including sympathy) behind social touch, either through experiencing or observing the social touch (Hertenstein, Keltner, App, Bulleit, & Jaskolka, 2006). Despite the deliberate nature of physical gestures, this behavior is emotionally driven and emotionally salient. Hence, comforting behavior can be described as a manifestation of lower-level aspects of empathy, such as the mirror neuron system. Moreover, a study found that higher levels of empathy might lead to more sympathetic social touch. During an experimental pain induction task in which couples were touching, levels of the observer's empathy predicted decreases in physical pain, suggesting that empathy is associated with social touch and pain reduction (Goldstein et al., 2016).

Social touch was shown to reduce distress and increase positive feelings, thus contributing to interpersonal emotion regulation (Goldstein et al., 2016; Goldstein, Weissman-Fogel, Dumas, & Shamay-Tsoory, 2018; Nummenmaa et al., 2016; Peled-Avron, Levy-Gigi, Richter-Levin, Korem, & Shamay-Tsoory, 2016). For example, in a pain perception study, hand-holding significantly reduced levels of experienced pain (Krahé, Drabek, Paloyelis, & Fotopoulou, 2016). Further, parents frequently use physical methods of soothing (hugging, patting, etc.) to effectively calm their distressed children (Cekaite & Kvist, 2017).

Collectively it appears that emotional empathy leads to increased regulation of emotion through associated mimicry, synchrony, and physical touch, which predicts reduction of distress and increase in emotion regulation.

Cognitive Empathy and Interpersonal Emotion Regulation

Cognitive empathy consists of understanding the thoughts and motivations of others, and involves active perspective-taking (spontaneous adopting of the psychological point of view of others) and theory of mind (having meta-representations of minds of others). Cognitive empathy is not always related to helping reduce others' distress, but often is. According to our own model, it entails three processes: representing mental states, attributing these states to others, and applying this knowledge to understand the behavior of others (Abu-Akel & Shamay-Tsoory, 2011). It can further be divided into affective and cognitive theory of mind-making inferences about cognitions versus emotions, respectively (Abu-Akel & Shamay-Tsoory, 2011; Eslinger, 1998).

Theory of mind and mentalizing is associated with a number of brain regions, forming a complex network (Abu-Akel & Shamay-Tsoory, 2011; Bodden et al., 2013). Overall, the neurological basis of theory of mind lies primarily in a network of the medial PFC, temporal lobe (particularly the temporo-parietal junction), and posterior superior temporal sulcus (pSTS) (Dulau, 2015; Shamay-Tsoory, 2011; Singer, 2006). The neural basis for affective or cognitive theory of mind overlaps, but is not identical.

Considering that cognitive empathy is important for understanding the perspective of others, it may consequently help understanding which emotion regulation strategy fits a specific person in a specific situation. Cognitive empathy can contribute to deliberately selecting an emotion regulation strategy for the target, and verbal and behavioral affirmations of understanding the target's emotions. Once an individual understands what another person is feeling or thinking, it is possible for them to verbalize this, either by stating that they understand or through verbally labeling the target's emotions. Hence, many studies focus on developing cognitive empathy measures by asking participants to describe or label the emotions or thoughts of another person (Bensalah, Caillies, & Anduze, 2016).

This verbalization of understanding has the potential to reduce distress, thus contributing to interpersonal emotion regulation. As mentioned, feeling understood activates the ventral striatum and middle insula, brain regions associated with reward, thus increasing pleasure and feelings of social connection, and, perhaps in this way, reducing distress. Further, labeling emotions can reduce their intensity. One study showed that affect labeling reduces activation of the limbic system following negative stimulus, and increases activation of the PFC, specifically the right ventrolateral PFC (Lieberman et al., 2007). Though there is no research specifically focused on the effects of interpersonal emotion labeling, emotion socialization literature shows that discussing the causes of emotions can lead to better regulation (Denham, Zoller, & Couchoud, 1994). Additionally, many types of therapy emphasize the importance of labeling emotions in reducing their negative effect (e.g., Bai & Yue, 2013). Thus, cognitive empathy can lead to verbalized understanding of the target's emotions, which in turn leads to reward and regulation.

Cognitive empathy might also include deliberately attempting to alter the emotions of the target. This attempt can include a variety of strategies, such as giving advice, providing alternative explanations (reappraisal), distracting, and encouraging. Cognitive empathy is a precursor to choosing a strategy, as one must understand the emotions of the target in order to try and change those (Zaki & Williams, 2013). One study found developmental differences in the interpersonal emotion regulation strategies suggested by children. Younger children relied primarily on distraction, suggesting that developmental and cognitive factors (such as theory of mind) play a role in strategy selection. Strategies that required understanding of the emotional states of others were more frequently used by older children, providing support for the theory that cognitive empathy is a prerequisite for interpersonal emotion regulation strategy selection (López-Pérez, Wilson, Dellaria, & Gummerum, 2016).

Neuroimaging studies further exemplify the relationship between cognitive empathy and strategy selection. Recent studies (e.g., Hallam et al., 2014) demonstrated activation of the ventromedial PFC in strategy selection for interpersonal emotion regulation. Hallam et al. (2014) found that all areas typically activated in intrapersonal emotion regulation are activated in the empathizer during interpersonal emotion regulation, with the addition of the left anterior temporal pole and the medial PFC, areas associated with theory of mind. Different sub-regions are differentially activated by the type of strategy chosen. Thus, neuroanatomical evidence implicates regions associated with cognitive empathy and theory of mind with strategy selection in interpersonal emotion regulation.

As different emotion regulation strategies are effective in different distressing situations, and an individual in distress may not be able to rationally choose an effective strategy, an outside observer might be more effective in selecting a strategy than the distressed individual (English, Lee, John, & Gross, 2017; Levy-Gigi & Shamay-Tsoory, 2017). Indeed, Levy-Gigi and Shamay-Tsoory (2017) found that regulation strategies chosen and applied by a partner were more effective at reducing distress than intrapersonal emotion regulation. Therefore, cognitive empathy can lead to more effective interpersonal emotion regulation through better selection of regulation strategy. There is initial empirical support for this theory, as a recent study found that higher levels of cognitive empathy (assessed by questionnaire) predicted successful interpersonal emotion regulation (Levy-Gigi & Shamay-Tsoory, 2017).

Conclusion

Much research examined empathy and emotion regulation separately, as they are both constructs of significant relevance to clinical psychology. Here we synthesize different lines of studies into an integrative model of empathy, examining the contributions of empathy to distress regulation. We provide a new, more dyadic perspective towards interpersonal emotion regulation, one that accounts for the multiple paths through which these constructs influence each other, as well as the neurological basis of those paths. This model broadens the conception of interpersonal emotion regulation to include the contribution of empathy and empathetic responses. Further research is needed to elucidate the mechanisms of these processes. Though extant research supports each of the steps outlined in this model, more research is needed to confirm the cyclical nature of these steps, how the neural mechanisms of both the target and the empathizer interact in a reciprocal way, and their neurological and physiological bases.

Understanding the mechanisms through which empathy plays a role in interpersonal emotion regulation is important due to the role of emotion regulation in mental health and the potential to harness these mechanisms to aid individuals in distress (Beauregard et al., 2004; Levy-Gigi & Shamay-Tsoory, 2017). With more research, we might gain a better understanding of how to make these interpersonal emotion regulation processes intentional, so that everyone might be better equipped to reduce distress and dysfunction in people close to them (Fig. 1).

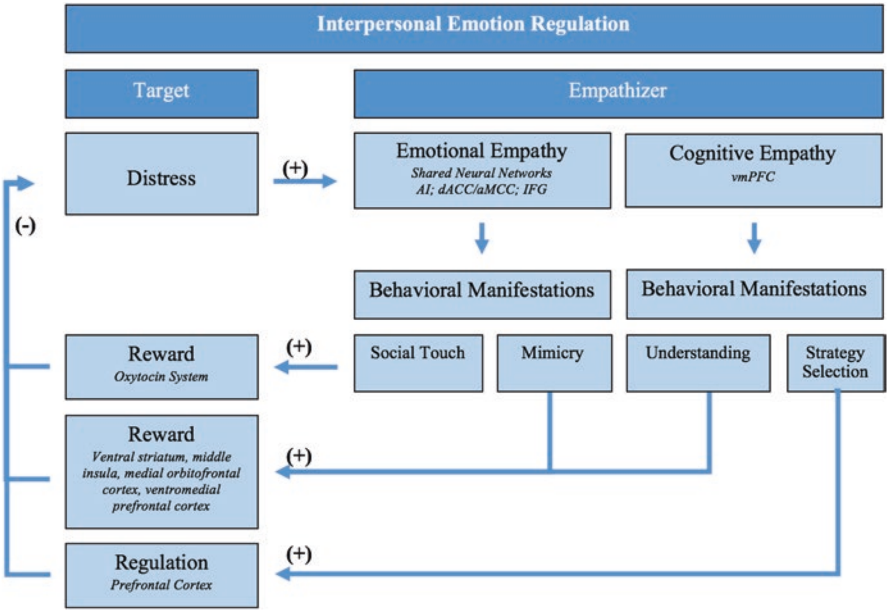


Fig. 1 The distress of the target evokes both emotional and cognitive empathy in the empathizer, which both uniquely contribute to the regulation of that distress. Emotional empathy is activated through shared neural networks centered in the anterior insula (AI), the dorsal-anterior/anterior-midcingulate cortex (dACC/aMCC), and inferior frontal gyrus (IFG). Cognitive empathy is activated through a network of regions centered in the ventromedial prefrontal cortex (vmPFC). Emotional empathy contributes to reduction of distress in two ways: mimicry, synchrony and social touch, both of which are prompted by the neural mechanisms underlying emotional empathy. Mimicry reduces the target's distress by activating reward through the ventral striatum, middle insula, medial orbitofrontal cortex, and ventromedial prefrontal cortex. Social touch similarly reduces the target's distress by activating reward through the oxytocin system. Cognitive empathy contributes to interpersonal emotion regulation through strategy selection—deliberately attempting to alter the emotional state of the target. This then leads to regulation in the target through activation of regulation techniques centered in the prefrontal cortex. In these ways, emotional and cognitive empathy play unique roles in interpersonal emotion regulation and differentially contribute to the reduction of distress in the target, which then, in turn, leads to reduction of empathetic arousal in the empathizer

References

Abu-Akel, A., & Shamay-Tsoory, S. (2011). Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia, 49*(11), 2971–2984.

Bai, X., & Yue, P. (2013). Affect labeling can reduce negative emotions: Evidences from autonomic nervous activity. *Acta Psychologica Sinica, 45*(7), 715–724.

Beauregard, M., Lévesque, J., & Paquette, V. (2004). Neural basis of conscious and voluntary self-regulation of emotion. In M. Beauregard (Ed.), *Advances in consciousness research. Consciousness, emotional self-regulation and the brain* (pp. 163–194). Amsterdam, Netherlands: John Benjamins Publishing Company.

Bensalah, L., Caillies, S., & Anduze, M. (2016). Links among cognitive empathy, theory of mind, and affective perspective taking by young children. *The Journal of Genetic Psychology: Research and Theory on Human Development, 177*(1), 17–31.

Berhardt, B., & Singer, T. (2012). The neural basis of empathy. *Annual Review of Neuroscience, 35*, 1–23.

Bodden, M. E., Kübler, D., Knake, S., Menzler, K., Heverhagen, J. T., Sommer, J., … Dodel, R. (2013). Comparing the neural correlates of affective and cognitive theory of mind using fMRI: Involvement of the basal ganglia in affective theory of mind. *Advances in Cognitive Psychology, 9*(1), 32–43.

Budell, L., Jackson, P., & Rainville, P. (2010). Brain responses to facial expressions of pain: Emotional or motor mirroring? *NeuroImage, 53*(1), 355–363.

Buhle, J. T., Silvers, J. A., Wager, T. D., Lopez, R., Onyemekwu, C., Kober, H., … Ochsner, K. N. (2014). Cognitive reappraisal of emotion: A meta-analysis of human neuroimaging studies. *Cerebral Cortex (New York, N.Y. : 1991), 24*(11), 2981–2990.

Cekaite, A., & Kvist, M. H. (2017). The comforting touch: Tactile intimacy and talk in managing children's distress. *Research on Language and Social Interaction, 50*(2), 109–127.

Colibazzi, T., Posner, J., Wang, Z., Gorman, D., Gerber, A., Yu, S., … Peterson, B. S. (2010). Neural systems subserving valence and arousal during the experience of induced emotions. *Emotion, 10*(3), 377–389.

Cuff, B. M. P., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: A review of the concept. *Emotion Review, 8*(2), 144–153.

de Waal, F. B. M. (2008). Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology, 59*, 279–300.

de Waal, F. B. M., & Preston, S. D. (2017). Mammalian empathy: Behavioural manifestations and neural basis. *Nature Reviews Neuroscience, 18*(8), 498–509.

Denham, S. A., Zoller, D., & Couchoud, E. A. (1994). Socialization of preschoolers' emotion understanding. *Developmental Psychology, 30*(6), 928–936.

Dimberg, U., Andréasson, P., & Thunberg, M. (2011). Emotional empathy and facial reactions to facial expressions. *Journal of Psychophysiology, 25*(1), 26–31.

Dulau, C. (2015). Introduction to social cognition. In B. Brochet (Ed.), *Neuropsychiatric symptoms of neurological disease. Neuropsychiatric symptoms of inflammatory demyelinating diseases* (pp. 181–194). Cham, Switzerland: Springer International Publishing.

Eisenberg, N., & Eggum, N. D. (2009). Empathic responding: Sympathy and personal distress. In J. Decety & W. Ickes (Eds.), *Social neuroscience. The social neuroscience of empathy* (pp. 71–83). Cambridge, MA: MIT Press.

English, T., Lee, I. A., John, O. P., & Gross, J. J. (2017). Emotion regulation strategy selection in daily life: The role of social context and goals. *Motivation and Emotion, 41*(2), 230–242.

Eres, R., Decety, J., Louis, W. R., & Molenberghs, P. (2015). Individual differences in local gray matter density are associated with differences in affective and cognitive empathy. *NeuroImage, 117*, 305–310.

Eslinger, P. J. (1998). Neurological and neuropsychological bases of empathy. *European Neurology, 39*(4), 193–199.

Etkin, A., Büchel, C., & Gross, J. J. (2015). The neural bases of emotion regulation. *Nature Reviews Neuroscience, 16*(11), 693–700.

Feldman, R. (2007a). Parent-infant synchrony: Biological foundations and developmental outcomes. *Current Directions in Psychological Science, 16*(6), 340–345.

Feldman, R. (2007b). Parent–infant synchrony and the construction of shared timing: Physiological precursors, developmental outcomes, and risk conditions. *Journal of Child Psychology and Psychiatry, 48*, 329–354.

Goldstein, P., Shamay-Tsoory, S. G., Yellinek, S., & Weissman-Fogel, I. (2016). Empathy predicts an experimental pain reduction during touch. *The Journal of Pain, 17*(10), 1049–1057.

Goldstein, P., Weissman-Fogel, I., Dumas, G., & Shamay-Tsoory, S. G. (2018). Brain-to-brain coupling during handholding is associated with pain reduction. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 115*(11), E2528–E2537.

Gonzalez-Liencres, C., Shamay-Tsoory, S. G., & Brüne, M. (2013). Towards a neuroscience of empathy: Ontogeny, phylogeny, brain mechanisms, context and psychopathology. *Neuroscience and Biobehavioral Reviews, 37*(8), 1537–1548.

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology, 85*(2), 348–362.

Grynberg, D., & Lopez-Perez, B. (2018). Facing others' misfortune: Personal distress mediates the association between maladaptive emotion regulation and social avoidance. *PLoS One, 13*(3), e0194248.

Hallam, G. P., Webb, T. L., Sheeran, P., Miles, E., Niven, K., Wilkinson, I. D., … Farrow, T. F. D. (2014). The neural correlates of regulating another person's emotions: An exploratory fMRI study. *Frontiers in Human Neuroscience, 8*, 376.

Hertenstein, M., & Weiss, S. (2011). *The handbook of touch*. New York, NY: Springer.

Hertenstein, M. J., Keltner, D., App, B., Bulleit, B. A., & Jaskolka, A. R. (2006). Touch communicates distinct emotions. *Emotion, 6*(3), 528–533.

Horn, A. B., & Maercker, A. (2016). Intra- and interpersonal emotion regulation and adjustment symptoms in couples: The role of co-brooding and co-reappraisal. *BMC Psychology, 4*, 51.

Jospe, K., Flöel, A., & Lavidor, M. (2018). The interaction between embodiment and empathy in facial expression recognition. *Social Cognitive and Affective Neuroscience, 13*(2), 203–215.

Krahé, C., Drabek, M. M., Paloyelis, Y., & Fotopoulou, A. (2016). Affective touch and attachment style modulate pain: A laser-evoked potentials study. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*, 20160009. https://doi.org/10.1098/rstb.2016.0009

Kuhn, S., Muller, B. C. N., van Baaren, R. B., Wietzker, A., Dijksterhuis, A., & Brass, M. (2010). Why do I like you when you behave like me? Neural mechanisms mediating positive consequences of observing someone being imitated. *Social Neuroscience, 12*, 1–9.

Kühn, S., Müller, B. C. N., van der Leij, A., Dijksterhuis, A., Brass, M., & van Baaren, R. B. (2011). Neural correlates of emotional synchrony. *Social Cognitive and Affective Neuroscience, 6*(3), 368–374.

Lamm, C., & Silani, G. (2013). Insights into collective emotions from the social neuroscience of empathy. In *Collective emotions: Perspectives from psychology, philosophy, and sociology* (pp. 63–77). Oxford University Press.

Levy-Gigi, E., & Shamay-Tsoory, S. G. (2017). Help me if you can: Evaluating the effectiveness of interpersonal compared to intrapersonal emotion regulation in reducing distress. *Journal of Behavior Therapy and Experimental Psychiatry, 55*, 33–40.

Lieberman, M. D., Eisenberger, N. I., Crockett, M. J., Tom, S. M., Pfeifer, J. H., & Way, B. M. (2007). Putting feelings into words: Affect labeling disrupts amygdala activity in response to affective stimuli. *Psychological Science, 18*(5), 421–428.

López-Pérez, B., Wilson, E. L., Dellaria, G., & Gummerum, M. (2016). Developmental differences in children's interpersonal emotion regulation. *Motivation and Emotion, 40*(5), 767–780.

Main, A., Walle, E. A., Kho, C., & Halpern, J. (2017). The interpersonal functions of empathy: A relational perspective. *Emotion Review, 9*(4), 358–366.

Moore, A., Gorodnitsky, I., & Pineda, J. (2012). EEG mu component responses to viewing emotional faces. *Behavioural Brain Research, 226*(1), 309–316.

Niven, K. (2017). The four key characteristics of interpersonal emotion regulation. *Current Opinion in Psychology, 17*, 89–93.

Nummenmaa, L., Tuominen, L., Dunbar, R., Hirvonen, J., Manninen, S., Arponen, E., … Sams, M. (2016). Social touch modulates endogenous μ-opioid system activity in humans. *NeuroImage, 138*, 242–247.

Peled-Avron, L., Levy-Gigi, E., Richter-Levin, G., Korem, N., & Shamay-Tsoory, S. G. (2016). The role of empathy in the neural responses to observed human social touch. *Cognitive, Affective, & Behavioral Neuroscience, 16*(5), 802–813.

Ramseyer, F., & Tschacher, W. (2011). Nonverbal synchrony in psychotherapy: Coordinated body movement reflects relationship quality and outcome. *Journal of Consulting and Clinical Psychology, 79*(3), 284–295.

Reeck, C., Ames, D. R., & Ochsner, K. N. (2016). The social regulation of emotion: An integrative, cross-disciplinary model. *Trends in Cognitive Sciences, 20*(1), 47–63.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*, 169–192.

Shamay-Tsoory, S. G. (2011). The neural bases for empathy. *The Neuroscientist, 17*(1), 18–24.

Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: A double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain: A Journal of Neurology, 132*(3), 617–627.

Shipman, K., Schneider, R., & Brown, A. (2004). Emotion dysregulation and psychopathology. In M. Beauregard (Ed.), *Advances in consciousness research. Consciousness, emotional self-regulation and the brain* (pp. 61–85). Amsterdam, Netherlands: John Benjamins Publishing Company.

Singer, T. (2006). The neuronal basis and ontogeny of empathy and mind reading: Review of literature and implications for future research. *Neuroscience and Biobehavioral Reviews, 30*(6), 855–863.

Singer, T., & Klimecki, O. (2014). Empathy and compassion. *Current Biology, 24*(18), 875–878.

Smith, A. (2006). Cognitive empathy and emotional empathy in human behavior and evolution. *The Psychological Record, 56*(1), 3–21.

Webb, T. L., Miles, E., & Sheeran, P. (2012). Dealing with feeling: A meta-analysis of the effectiveness of strategies derived from the process model of emotion regulation. *Psychological Bulletin, 138*(4), 775–808.

Yang, C.-Y., Decety, J., Lee, S., Chen, C., & Cheng, Y. (2009). Gender differences in the my rhythm during empathy for pain: An electroencephalographic study. *Brain Research, 1251*, 176–184.

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological Science, 19*, 399–404. https://doi.org/10.1111/j.1467-9280.2008.02099.x

Zaki, J., & Williams, W. C. (2013). Interpersonal emotion regulation. *Emotion, 13*(5), 803–810.

The Role of Mentalizing in Communication Behaviors

Jacob Parelman, Bruce Doré, and Emily B. Falk

Introduction

How does an idea move from one mind to another? Communication between people shapes our perceptions of the world and the behaviors we choose to enact. Our ability to navigate complex social relationships developed as our ancestors began forming larger and more complex groups (Dunbar, 1998). As modern humans, we have inherited remarkable social abilities, which allow us to effectively share knowledge, learn from others, and shape our behaviors around their experiences (Bandura, 1962). These communication skills, in part, rely on a process of mentalizing—thinking about other people's mental states (Frith & Frith, 2003). Within the brain, a mentalizing network including regions such as the medial prefrontal cortex (mPFC), temporal-parietal junction (TPJ), posterior cingulate cortex (PCC), precuneus (PC), and posterior superior temporal sulcus (pSTS) (Frith & Frith, 2006; Mitchell, 2009; Saxe, Carey, & Kanwisher, 2004; Spunt & Lieberman, 2012; Van Overwalle, 2009) occupies greater relative space in humans compared to other species (Bradbury, 2005; Nimchinsky et al., 1999) and facilitates social communication, among other tasks (Cacioppo & Cacioppo, 2013).

Here we describe research that underscores the role of mentalizing in successful communication, drawing on research from interpersonal and mass communication, economic decision-making, and social neuroscience. Specifically, these fields together highlight the critical role that mentalizing plays in guiding information

J. Parelman (✉) · B. Doré
Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, USA
e-mail: jparelman@falklab.org

E. B. Falk
Department of Psychology, Annenberg School for Communication, Philadelphia, PA, USA

Marketing Department, University of Pennsylvania, Philadelphia, PA, USA

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_30

sharing decisions, facilitating peer influence and behavior change, and promoting shared understanding across communicators and receivers.

Mentalizing in Communication

To understand how mentalizing is used during communication, we first describe its role in facilitating information sharing and message reception. Individuals often flow between acting as communicators and receivers of information over the course of an interaction; however, these two roles can be described as distinct in how they take advantage of our ability to infer others' knowledge and intentions. In addition to exploring the process of mentalizing in these distinct roles, we also review how regions of the mentalizing network are recruited both when individuals take on the role of communicator and the role of receiver.

Communicators

"Know your audience." This advice highlights that considering the knowledge, thoughts or intentions of one's audience are crucial for effective communication. Constructing an effective message involves accurately representing others' minds, and this process can be very effortful and calculated (Frenzen & Nakamoto, 1993), or automatic and effortless—shaping not only what we say, but how we say it (Berger, 2014). As an example, imagine you're approached in a park near your house by a stranger asking for directions, but they clearly do not speak your native language fluently. Immediately, you begin a process of inferring what knowledge this stranger has of the city, and how best to help them. How you choose to direct this stranger, and in what manner, will no doubt be based on your inferences (Kingsbury, 1968; Krauss & Fussell, 1991), and how helpful you are to this person will depend on how accurately you are able to represent their knowledge and goals. This interaction illustrates one way that mentalizing contributes to the ultimate success of social decision-making and communication: supporting social inferences.

How Is Mentalizing Utilized for Generating and Sharing Information?

Our example in the park illustrates how mentalizing facilitates social inference and message production: initial impressions are used to represent what information a target might need for the goals we perceive them to have. These representations are not static however, and mentalizing makes it possible to continuously update knowledge about others and what information to share with them. Now imagine learning after your initial advice that the stranger in the park is, in fact, from the city. How might this new information shape your next instruction? Recipient design theory

(Sacks, Schegloff, & Jefferson, 1978) posits that mentalizing is used to guide communication both before an interaction, through initial impressions, and as feedback and new information about a target is incorporated into representations of their knowledge and mental state. In this section, we use the recipient design theory as a framework to describe the role of mentalizing in message production and communication.

Beginning with initial impression formation, communicators use a variety of information sources to produce messages appropriate for the perceived needs of specific individuals. The physical location of a target and their perspective (Dumontheil, Küster, Apperly, & Blakemore, 2010; Keysar, Barr, Balin, & Brauner, 2000), target identity (Galati & Brennan, 2010), shared history or knowledge (Fussell & Krauss, 1992), and other factors shape message content. Experimental communication tasks, like the Tacit Communication Game (TCG) (De Ruiter, Noordzij, Newman-Norlund, Hagoort, & Toni, 2007), are one way that scientists have studied the effects of these information streams on message content and communicator decision-making more broadly. In the tacit communication game, communicators are asked to guide a partner, or receiver, to a hidden goal on a 3 × 3 grid using only vertical and horizontal movements. As part of their instructions, communicators are given freedom to move in any direction, at any speed, and with as many moves as they would like, thus providing variability in communicative strategy. The identity of the receiver may be varied in the tacit communication game, and it is this manipulation in which message tailoring can be experimentally controlled and investigated. For example, in one tacit communication game study, adult communicators were made to believe that they were either guiding another adult or a child to their goal (Newman-Norlund et al., 2009). This simple alteration dramatically changed the strategy that communicators used, such that instructive actions were deliberately slower and more repetitive near the target when communicators thought they were playing with a child. This study, and others like it, provides clear evidence that communicative decision-making is affected by the inferences that communicators make about the knowledge and abilities of their intended audience.

The tacit communication game has also helped to link the brain's mentalizing network to recipient design and message tailoring. Patients with damage to the vmPFC, a region often implicated in mentalizing (Atique, Erb, Gharabaghi, Grodd, & Anders, 2011; Schiller, Freeman, Mitchell, Uleman, & Phelps, 2009), show efforts to convey useful instructions to targets while playing the tacit communication game, but fail to make communicative accommodations for children and adults, respectively (Stolk, D'Imperio, di Pellegrino, & Toni, 2015). Damage to this region seemed to impact communicators' ability to modify their instructions for different receivers. This lesion study, in conjunction with other correlational neuroimaging studies that associate the mentalizing network with message tailoring (Kuhlen, Bogler, Brennan, & Haynes, 2017; Noordzij et al., 2009; Vanlangendonck, Willems, & Hagoort, 2018), suggests that mentalizing is an important feature of recipient design and the process of message formation and delivery.

Recipient design does not end with communicators' initial impressions—the theory also contends that communicators incorporate feedback from their target over the course of their interactions and as new information is learned (Newman-Norlund et al., 2009). Indeed, people can be remarkably sensitive to their communication partners—quickly changing course or doubling down at the slightest wince or smile, boo or cheer. Here also, mentalizing is involved in communication strategy revision (Bögels et al., 2015). For example, communicators in the tacit communication game show greater engagement of the TPJ when receiving feedback from their receiver, which in turn relates to changes in instruction (Bögels et al.,2015). Activation in the STS, rIPL, and PCC is also associated with impression updating (Mende-Siedlecki, Cai, & Todorov, 2013), and the tracking of individual characteristics over time (Cloutier, Gabrieli, O'Young, & Ambady, 2011; Ma et al., 2012). Additionally, some of these same regions (STS, TPJ) are implicated in tracking relevant characteristics of other social agents during economic decision-making tasks (Behrens, Hunt, Woolrich, & Rushworth, 2008). This research provides evidence that people continuously incorporate feedback and new information into mental models of interacting partners in the context of active communication, in more basic forms of decision-making and behavior revision, and, importantly, activate regions of the brain's mentalizing system to guide decisions.

How Does Mentalizing Lead to Successful Communication?

Mentalizing has an important role in providing a knowledge base for communicators to inform what information to share, but successful communication may hinge on whether a communicator can accurately represent the mental states of their audience (e.g., states of knowledge or belief). In a classic series of studies (Fussell and Krauss 1989), participants provided written descriptions of several shapes for either themselves at a later time, their friend, or a complete stranger. Friends and participants themselves performed significantly better than strangers using these written descriptions, a result which the authors contend is driven by communicators using language that is informed by their and their friends' "common-ground" (shared knowledge or beliefs (Clark & Murphy 1982)). Further, sharing an experience (even with an unknown target) provides enough common-ground for communicators to draw on when constructing more effective instruction (Traxler & Gernsbacher, 1993), a result that is consistent with the idea that the success of mentalizing is increased by a better understanding of a social target, and that this in turn facilitates more successful communication.

Applied research also finds that the success of communication is related to mentalizing processes within individuals. For example, research on how individuals successfully propagate information indicates that brain activity in key parts of the mentalizing system, including the TPJ, dmPFC, precuneus, and ventral-dorsal striatum, is more engaged for content that individuals go on to positively endorse and enthusiastically share (Falk, Morelli, Welborn, Dambacher, & Lieberman, 2013). Additionally, product ads that more actively engage the TPJ and dmPFC were also associated with more use of social appeals when participants promoted the same product (O'Donnell, Falk, & Lieberman, 2015).

Mentalizing also contributes to whether or not people share information with others in the first place. Indeed, neuroscience research shows that the spread of information may begin with simple social inferences (e.g., whether others will find information valuable or useful) on the part of individuals (Falk & Scholz, 2018). For example, one recent study found that when participants made decisions about sharing health news articles, activity in the TPJ, dmPFC, and PCC predicted their subsequent decisions to share the content (Baek et al., 2017). Further, those articles that elicited greater activity in the mentalizing network across participants also were shared more by a larger and separate population of news readers (Scholz et al., 2017) who may have also engaged in mentalizing as part of their communicative decision-making.

Finally, individual differences in the extent to which communicators recruit the TPJ (Falk et al., 2013) and mPFC (Dietvorst et al., 2009) track with their abilities as communicators and persuaders; salespeople who could acquire and maintain more profitable accounts also scored higher on a variety of mentalizing related skills like perspective taking, anticipating the needs of clients, detecting nonverbal cues, and shaping the course of the sales interactions (Dietvorst et al., 2009). These same high performing salespeople also showed greater activity in the mPFC during a mentalizing task compared to low performing salespeople. This "salesperson effect" (Falk et al., 2013), or greater tendency for more effective communicators to engage areas of the mentalizing system, parallels research showing that individuals who engage in greater mentalizing also tend to express more socially adaptive behaviors like cooperating more (Krach et al., 2009; Paal & Bereczkei, 2007; Ridinger & McBride, 2017), being more inclusive (Masten, Morelli, & Eisenberger, 2011), writing more persuasively to different audiences (Rubin & Rafoth, 1986), and more effectively negotiating (Galinsky et al., 2008).

The evidence, that mentalizing—and more specifically representing the mental states of communicative targets—facilitates communicative decisions and abilities, converges with a broader literature in neuroeconomics. This research has consistently found that individuals consider the mental states of others in order to guide their behavior, and that individuals with social deficits often perform poorly when making social decisions (Sally & Hill, 2006). Areas of the mentalizing network are frequently engaged when people play strategic games that require them to understand and predict the behavior of another player before making a move. For example, regions of the TPJ and ACC are both actively engaged in predicting the behavior of other people in competitive card-games or tasks (Carter, Bowling, Reeck, & Huettel, 2012; Gallagher, Jack, Roepstorff, & Frith, 2002), the ventral and dorsal mPFC, the pSTS, and PCC are all involved in tracking information about the beliefs of opponents in competitive tasks (Hampton, Bossaerts, & O'Doherty, 2008; Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004), and the mPFC is activated by considering how cooperative others are, as well as how cooperative one feels like being during such tasks (McCabe, Houser, Ryan, Smith, & Trouard, 2001). Complementing these findings, individuals with autism spectrum disorder, a population characterized by social deficits (Lombardo, Barnes, Wheelwright, & Baron-Cohen, 2007), not only fail to incorporate social inferences in economic decision-making tasks (Sanfey, 2007), but also show reduced engagement of the rTPJ compared to control participants when making socially relevant inferences (Lombardo et al., 2011).

Overall, research from social psychology, communication, and neuroscience indicates that mentalizing impacts how people communicate and interact with others. This growing body of evidence suggests that neural pathways implicated in mentalizing can predict how successful a message is in reaching an audience (Scholz et al., 2017), and how successful individuals are in convincing others (Dietvorst et al., 2009; Falk et al., 2013). Such research falls into a broader area of science that finds mentalizing and the mentalizing network as necessary for decision-making in social contexts.

Receivers

Reviewing how communicators use perspective taking to transmit ideas and persuade others considers only half of our story: listeners are at the other end of these exchanges. In this section, we explore evidence that information receivers also use their mentalizing skills to evaluate the content of messages and form preferences (Falk & Scholz, 2018). Again, findings from social psychology, economics, communication, and neuroscience provide parallel insights, suggesting that mentalizing is a general process involved in successful social decision-making across communicators and receivers.

How Does Mentalizing Help to Understand Communicated Information?

To understand information from other people, a receiver may need to consider the goals or intentions of the communicator. That is, the meaning of a gesture or comment can be affected by knowledge of the person (or entity) communicating it. Interactions can hinge on such an understanding (e.g., an inside joke or misread nod), and so the success of a message can rely not only on the message itself but on how a receiver understands the context inherent to mental states of the communicator.

The characteristics of a communicator can have a significant effect on how message receivers process and value information. These mediating factors, or source effects, are a topic of extensive research in social psychology and both basic and applied research in communication and consumer behavior (see Wilson & Sherrell, 1993 for review). Source effects like the credibility, expertise, trustworthiness (Petty, Cacioppo & Heesacker, 1984; Kang & Herr, 2006; Kumkale, Albarracin, & Seignourel, 2010; Sternthal, Phillips, & Dholakia, 1978), attractiveness (Chaiken, 1979), and ideological similarity of a communicator (Silvia, 2005; Woodside & Davenport, 1974) all have long histories of positive effects on message processing and attitudes or behavior change. Evidence from neuroscience further indicates that mentalizing processes are involved in these source effects. For instance, objects associated with attractive or high expertise celebrities are not only valued more by observers but also elicit greater activation in the dmPFC (Klucharev, Smidts, & Fernández, 2008); high status individuals elicit greater activation in the dmPFC, PC, and rTPJ when others view their faces (Zerubavel, Bearman, Weber, & Ochsner, 2015); and source identity cues like group affiliation (Stallen, Smidts, & Sanfey, 2013), race (Cikara & Van Bavel, 2014; Ito & Bartholow, 2009), and even religion (Bruneau, Dufour, & Saxe, 2012) relate to activation in areas of the mentalizing network when information is evaluated by receivers.

Do Receivers Vary in Their Sensitivity to Social Information?

Many of our decisions, whether it's what news to read (Hermida, Fletcher, Korell, & Logan, 2012), food to eat (Zhang, Ye, Law, & Li, 2010), or even medical choices (Frost & Massagli, 2008), involve the consideration and incorporation of social feedback. Even with anonymous peers, mentalizing and social comparison still influence decision-making (Cascio et al., 2015a; Klucharev, Hytönen, Rijpkema, Smidts, & Fernández, 2009). Likewise, while individuals are generally attentive to deviations from group recommendations during decision-making tasks, individuals' reactivity in the TPJ tracks with sensitivity to peer feedback, such that those individuals who show greater activation in the TPJ when viewing the opinions of a group are also more likely to update their opinions to fall in line with the group (Cascio et al., 2015b). Interestingly, such results also vary with individuals' social network structure (O'Donnell, Bayer, Cascio, & Falk, 2017), indicating that one's social environment may also impact the neural processes that give rise to conformity. These findings and similar research (Welborn et al., 2016) suggest that individuals may be differentially influenced by normative messages during consumer decisions, and that such variability may be explained by both environmental and neurobiological factors like social network structure and mentalizing network sensitivity.

Mentalizing, Sharing, and Interactive Information Transfer

Multiple lines of evidence converge to show that mentalizing is an important process both for communicators choosing how and what to share, and for receivers determining whether or not information is persuasive. Although these lines of inquiry address mentalizing in these two communicative roles, they don't address the process of information transfer itself. Given that communication necessarily involves two or more agents interacting, studying these roles in isolation does not completely encompass the processes involved. In this next section, we review research that indicates that the phenomenon of information transfer itself is supported by synchrony between people's mentalizing networks.

Inter-subject correlation (ISC), an analysis technique which measures the extent of shared neural processing between two or more individuals, has driven the understanding of the processes involved in information transfer and experience sharing. As part of this analysis method, either the spatial pattern of brain activity or (more commonly) the time-course of activation of two or more individuals' brains are compared for similarity as information is presented (Hasson, Nir, Levy, Fuhrmann, & Malach, 2004). The method is generally model free, which makes it particularly well suited for understanding how individuals similarly process and represent naturalistic stimuli (e.g., movies and written stories) or synchronize during realistic interpersonal interactions. ISC research has revealed that neural coupling occurs in areas of the brain responsible for basic perception (Silbert, Honey, Simony, Poeppel, & Hasson, 2014), the value system (Zadbood, Chen, Leong, Norman, & Hasson, 2017), and that during face-to-face interactions areas of the mentalizing network (rTPJ) show increased similarity between partners (Tang et al., 2016).

How Does Neural Synchrony Facilitate Communication?

As individuals interact, a complex process of mimicry and synchrony occurs in conversation topic and language use (Doré & Morris, 2018), prosodic cues (Lee et al., 2010), body position (Cappella, 1997), and even physiology (Mønster, Håkonsson, Eskildsen, & Wallot, 2016). Such coupling between individuals is thought to facilitate the transmission of information (Falk & Scholz, 2018), with the brains of two individuals sharing how information is represented both perceptually (Chen et al., 2017) and cognitively (Parkinson, Kleinbaum, & Wheatley, 2018). For example, in research by Stephens, Silbert, and Hasson (2010), speakers were instructed to tell a personal story while inside the MRI, and this story was then played to another listener while their brain was also scanned. Results from the study indicated that auditory processing areas, as well as the mPFC, dlPFC, striatum, precuneus, and TPJ were all significantly coupled between speaker-listener pairs, and importantly, that the extent to which speaker-listener brain signal was coupled in these areas was predictive of how successfully the listener could recall the speaker's story. To establish that the coupling–comprehension relationship was not driven by low-level linguistic or auditory features, the authors also showed that the relationship did not hold when speakers told stories in a language that the listener did not comprehend.

Beyond temporal synchrony, successful information transfer also evokes patterns of brain activity across speakers and listeners that are highly spatially similar (Zadbood et al., 2017). In one study, speakers described scenes from two television shows to listeners. Speakers and listeners showed significant spatial correlation in the precuneus, PCC, and mPFC, and the amount of speaker-listener pattern correlation in these regions was predictive of successful memory of the spoken information by listeners. Together with the results of Stephens et al. (2010), these results suggest that socially mediated information transfer depends on the coupling of neural signal over space and time in brain regions responsible not only for perception, but also higher order brain areas including the precuneus, mPFC, dmPFC, and TPJ.

How Do the Brains of Audiences Synchronize to Messages?

Neural synchrony also occurs between larger groups of individuals, and not just in cases of direct interpersonal communication. As audiences interact with messages, the extent to which a message is successful is also associated with the extent to which the neural signal between individuals in the group follow a similar pattern (Hasson et al., 2004; Schmälzle, Häcker, Honey, & Hasson, 2015). In these investigations, stimulus driven activity in the visual and auditory cortices are often correlated, but higher order regions of the mentalizing network such as the STS, mPFC, and TPJ are also correlated between observers. For example, Schmälzle et al. (2015) found that correlated activity in the TPJ and mPFC in response to political speeches was associated with the speeches being evaluated as stronger rather than weaker, suggesting that successful speeches result in shared processing of social information in the minds of listeners.

Overall these lines of research highlight the importance of understanding how individuals and groups interact as they engage in shared processing of information. This area of research still has much to explore, but has already started to reveal the importance of mentalizing for understanding socially mediated communication.

Future Directions

Beyond building an initial basic science model of the neuroscience of successful communication, it is also important to identify contextual factors that influence or moderate the effects of mentalizing on information sharing and persuasion. Two such contextual factors, intergroup bias and mediating technology, are particularly relevant to practitioners and researchers focusing on modern social life, and may be fruitful topics for researchers in this area to explore.

Communication Breakdown: Mentalizing and Intergroup Bias

Social conflict is common across the globe, and understanding how group biases impact communication is of great importance for improving discourse between groups and promoting social understanding. A broad literature in social psychology and neuroscience indicates that group identity affects decision-making (Bodenhausen, 1988; Bruneau & Saxe, 2010; Cikara, Botvinick, & Fiske, 2011) and social perception (Van Bavel, Packer, & Cunningham, 2011). Such effects are so salient, in fact, that even arbitrary, experimentally constructed, groups can powerfully shape responses to in- and out-group members (Brewer, 1979; Judd & Park, 1988; Taijfel, 1970; Van Bavel, Packer, & Cunningham, 2008). Mentalizing is also affected by group bias—people are more conservative in their attribution of mental capacities when observing the faces of out-group members (Hackel, Looser, & Van Bavel, 2014), and even show reduced empathic response in the mentalizing and pain networks during exposure to the pain or suffering of out-group members (Cikara, Bruneau, & Saxe, 2011).

One informative direction for this research area could be in exploring how mentalizing, or a lack of mentalizing, toward out-group members can lead to reduced civility and fairness in communication (Galinsky & Moskowitz, 2000), as inferences about a target seem to have a dramatic impact on how a communicator shapes the content of a message (Noordzij et al., 2009). A question worthy of greater attention is whether failed communication between members of opposing groups results from a lack of perspective taking, or engagement in inaccurate perspective taking based on false stereotypes. Further, only a handful of neuroimaging studies have asked how communicators update their inferences about targets from their in- versus out-group (Bögels et al., 2015; Freeman, Schiller, Rule, & Ambady, 2010). Contributions in this area could help explain how stereotypes or false assumptions may be corrected when people engage in conversation. Such work could build our scientific understanding of how people shape their statements when confronting others they staunchly disagree with.

Mediated Mentalizing: How Distance Shapes Communication

A growing proportion of social interactions occur in a manner that is mediated by technology, making it especially important for researchers to understand how technologies affect communication. People sometimes find it difficult to interpret the meaning or intention of emails or texts, and empirical evidence indicates that people are more likely to misjudge the intentions of others over computer-mediated, versus face-to-face, communications (Kato & Akahori, 2005). Technologically mediated communication by its nature reduces the amount of contextual information, like eye-gaze or gesture, available to a recipient (Sproull & Kiesler, 1986). This is important because such secondary communicative information can improve interpersonal understanding (Kiesler, Siegel, & McGuire, 1984) and cooperation (Tang et al., 2016). Recent fNIRS neuroimaging studies suggest that the positive effects of secondary communicative information may be related to greater mentalizing in response to richer information, in that activation in the mentalizing network—especially the TPJ—is greater when individuals interact face-to-face as compared to when they are separated by physical barriers (Jahng, Kralik, Hwang, & Jeong, 2017; Tang et al., 2016).

An important feature of online communication is that it can be spatially, temporally, and socially distant—depending on the platform, other individuals may not be in immediate proximity, may communicate asynchronously, and may or may not be perceived as immediate social entities (Norman, Tjomsland, & Huegel, 2016). In the brain, the dmPFC, a region in the mentalizing network, is more active when individuals evaluate information that is perceived as more psychologically distant (Baetens, Ma, Steen, & Van Overwalle, 2014). Combining this finding with the noted role of mentalizing in communication reveals a set of interesting questions. Namely, future research may ask whether online communications with different affordances differ as a function of how they affect mentalizing and activation in the mentalizing network. Evidence already suggests that psychological distance and modality do impact cognitive processes, like how communicative information is attended to and remembered (Amit et al., 2019; Amit, Algom, & Trope, 2009), and how content is assessed and valued (Henderson, Wakslak, Fujita, & Rohrbach, 2011). Similar research could help to disambiguate whether spatial, temporal, or hypothetical distance have similar or independent effects on mentalizing as well. For instance, would a communication medium like text messaging have varying effects on an individual's ability to infer the mental state of their partner or fall into neural synchrony if the time between sending and receiving messages was shortened, thus reducing temporal distance while maintaining spatial distance? Although it is difficult to manipulate some of these factors within the constraints of fMRI (e.g., spatial distance), methods such as fNIRS and EEG may offer more flexibility for naturalistic assessment (Vettel et al., 2019). Additionally, this research has the exciting potential to catalyze cross-discipline collaboration, further linking communication and neuroscience with related fields like linguistics and computer science.

Conclusion

The complexity and effectiveness of human communication is perhaps one key ingredient to our success as a species. Human communication is strongly facilitated by our ability to accurately infer what information should be shared with others and how to interpret information that is shared with us. The mentalizing system is implicated in a broad set of behaviors related to communication and decision-making, and this network is engaged when we both automatically and effortfully represent the mental states of our communicative partners. When acting as communicators, the mentalizing system facilitates our ability to infer the mental states of our audience in order to tailor how and what we say, and when acting as receivers, the system is engaged in relation to our attempts to understand messages and the intentions behind them. Neural synchrony between communicators and receivers also facilitates the flow of information between them. Environmental and situational factors impact the association between the mentalizing system and communicative decision-making, and it is these factors where some of the greatest promise for this area of research can be found. By linking issues in communication to neurobiological correlates, we will be able to better understand how the world we make and the world we live in impact our ability to share and connect with others at the most basic level.

References

Amit, E., Algom, D., & Trope, Y. (2009). Distance-dependent processing of pictures and words. *Journal of Experimental Psychology: General, 138*(3), 400.

Amit, E., Rim, S., Halbeisen, G., Priva, U. C., Stephan, E., & Trope, Y. (2019). Distance-dependent memory for pictures and words. *Journal of Memory and Language, 105*, 119–130.

Atique, B., Erb, M., Gharabaghi, A., Grodd, W., & Anders, S. (2011). Task-specific activity and connectivity within the mentalizing network during emotion and intention mentalizing. *NeuroImage, 55*(4), 1899–1911.

Baek, Elisa C., Christin Scholz, Matthew Brook O'Donnell, and Emily B. Falk. (2017). The Value of Sharing Information: A Neural Account of Information Transmission. *Psychological Science 28*(7), 851–61.

Baetens, K., Ma, N., Steen, J., & Van Overwalle, F. (2014). Involvement of the mentalizing network in social and non-social high construal. *Social Cognitive and Affective Neuroscience, 9*(6), 817–824.

Bandura, A. (1962). Social learning through imitation. In M. R. Jones (Ed.), *Nebraska Symposium on Motivation* (pp. 211–274). Lincoln, NE: University of Nebraska Press.

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature, 456*(7219), 245–249.

Berger, J. (2014). Word of mouth and interpersonal communication: A review and directions for future research. *Journal of Consumer Psychology, 24*(4), 586–607.

Bodenhausen, G. V. (1988). Stereotypic biases in social decision making and memory: Testing process models of stereotype use. *Journal of Personality and Social Psychology, 55*(5), 726–737.

Bögels, S., Barr, D. J., Garrod, S., & Kessler, K. (2015). Conversational interaction in the scanner: Mentalizing during language processing as revealed by MEG. *Cerebral Cortex, 25*(9), 3219–3234.

Bradbury, J. (2005). Molecular insights into human brain evolution. *PLoS Biology, 3*(3), e50.

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin, 86*(2), 307.

Bruneau, E. G., Dufour, N., & Saxe, R. (2012). Social cognition in members of conflict groups: Behavioural and neural responses in Arabs, Israelis and South Americans to each other's misfortunes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 367*(1589), 717–730.

Bruneau, E. G., & Saxe, R. (2010). Attitudes towards the outgroup are predicted by activity in the precuneus in Arabs and Israelis. *NeuroImage, 52*(4), 1704–1711.

Cacioppo, J. T., & Cacioppo, S. (2013). Social neuroscience. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 8*(6), 667–669.

Cappella, J. N. (1997). Behavioral and judged coordination in adult informal social interactions: Vocal and kinesic indicators. *Journal of Personality and Social Psychology, 72*(1), 119.

Carter, R. M., Bowling, D. L., Reeck, C., & Huettel, S. A. (2012). A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science, 337*(6090), 109–111.

Cascio, C. N., Carp, J., O'Donnell, M. B., Tinney, F. J., Jr., Bingham, C. R., Shope, J. T., … Falk, E. B. (2015a). Buffering social influence: Neural correlates of response inhibition predict driving safety in the presence of a peer. *Journal of Cognitive Neuroscience, 27*(1), 83–95.

Cascio, C. N., O'Donnell, M. B., Bayer, J., Tinney, F. J., & Falk, E. B. (2015b). Neural correlates of susceptibility to group opinions in online word-of-mouth recommendations. *JMR, Journal of Marketing Research, 52*(4), 559–575.

Chaiken, S. (1979). Communicator physical attractiveness and persuasion. *Journal of Personality and Social Psychology, 37*(8), 1387.

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience, 20*(1), 115–125.

Cikara, M., Botvinick, M. M., & Fiske, S. T. (2011). Us versus them: Social identity shapes neural responses to intergroup competition and harm. *Psychological Science, 22*(3), 306–313.

Cikara, M., Bruneau, E. G., & Saxe, R. R. (2011). Us and them: Intergroup failures of empathy. *Current Directions in Psychological Science, 20*(3), 149–153.

Cikara, M., & Van Bavel, J. J. (2014). The neuroscience of intergroup relations: An integrative review. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 9*(3), 245–274.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. *In Advances in psychology* , 9, 287–299, North-Holland.

Cloutier, J., Gabrieli, J. D. E., O'Young, D., & Ambady, N. (2011). An fMRI study of violations of social expectations: When people are not who we expect them to be. *NeuroImage, 57*(2), 583–588.

De Ruiter, J. P., Noordzij, M., Newman-Norlund, S., Hagoort, P., & Toni, I. (2007). On the origin of intentions. In P. Haggard, Y. Rossetti, & M. Kawato (Eds.), *Attention and performance XXII: Sensori motor foundation of higher cognition* (pp. 593–610). Oxford, England: Oxford University Press.

Dietvorst, R. C., Verbeke, W. J., Bagozzi, R. P., Yoon, C., Smits, M., & Van Der Lugt, A. (2009). A sales force–specific theory-of-mind scale: Tests of its validity by classical methods and functional magnetic resonance imaging. *Journal of Marketing Research, 46*(5), 653–668.

Doré, B. P., & Morris, R. R. (2018). Linguistic synchrony predicts the immediate and lasting impact of text-based emotional support. *Psychological Science, 29*(10), 1716–1723.

Dumontheil, I., Küster, O., Apperly, I. A., & Blakemore, S.-J. (2010). Taking perspective into account in a communicative task. *NeuroImage, 52*(4), 1574–1583.

Dunbar, R. I. (1998). The social brain hypothesis. *Brain: A Journal of Neurology, 9*(10), 178–190.

Falk, E., & Scholz, C. (2018). Persuasion, influence, and value: Perspectives from communication and social neuroscience. *Annual Review of Psychology, 69*, 329–356.

Falk, E. B., Morelli, S. A., Welborn, B. L., Dambacher, K., & Lieberman, M. D. (2013). Creating buzz: The neural correlates of effective message propagation. *Psychological Science, 24*(7), 1234–1242.

Freeman, J. B., Schiller, D., Rule, N. O., & Ambady, N. (2010). The neural origins of superficial and individuated judgments about ingroup and outgroup members. *Human Brain Mapping, 31*(1), 150–159.

Frenzen, J., & Nakamoto, K. (1993). Structure, cooperation, and the flow of market information. *The Journal of Consumer Research, 20*(3), 360–375.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50*(4), 531–534.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 358*(1431), 459–473.

Frost, J. H., & Massagli, M. P. (2008). Social uses of personal health information within PatientsLikeMe, an online patient community: What can happen when patients have access to one another's data. *Journal of Medical Internet Research, 10*(3), e15.

Fussell, Susan R., and Robert M. Krauss. (1989). Understanding Friends and Strangers: The Effects of Audience Design on Message Comprehension. *European Journal of Social Psychology 19*(6), 509–25.

Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology, 62*(3), 378–391.

Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language, 62*(1), 35–51.

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology, 78*(4), 708.

Galinsky, Adam D., William W. Maddux, Debra Gilin, and Judith B. White. (2008). Why It Pays to Get inside the Head of Your Opponent: The Differential Effects of Perspective Taking and Empathy in Negotiations. *Psychological Science 19*(4), 378–84.

Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage, 16*(3 Pt 1), 814–821.

Hackel, L. M., Looser, C. E., & Van Bavel, J. J. (2014). Group membership alters the threshold for mind perception: The role of social identity, collective identification, and intergroup threat. *Journal of Experimental Social Psychology, 52*, 15–23.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences of the United States of America, 105*(18), 6741–6746.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science, 303*(5664), 1634–1640.

Henderson, M. D., Wakslak, C. J., Fujita, K., & Rohrbach, J. (2011). Construal level theory and spatial distance. *Social Psychology, 42*(3), 165–173.

Hermida, A., Fletcher, F., Korell, D., & Logan, D. (2012). Share, like, recommend. *Journalism Studies, 13*(5–6), 815–824.

Ito, T. A., & Bartholow, B. D. (2009). The neural correlates of race. *Trends in Cognitive Sciences, 13*(12), 524–531.

Jahng, J., Kralik, J. D., Hwang, D.-U., & Jeong, J. (2017). Neural dynamics of two players when using nonverbal cues to gauge intentions to cooperate during the Prisoner's Dilemma Game. *NeuroImage, 157*, 263–274.

Judd, C. M., & Park, B. (1988). Out-group homogeneity: Judgments of variability at the individual and group levels. *Journal of Personality and Social Psychology, 54*(5), 778.

Kang, Y.-S., & Herr, P. M. (2006). Beauty and the beholder: Toward an integrative model of communication source effects. *The Journal of Consumer Research, 33*(1), 123–130.

Kato, Y., & Akahori, K. (2005). Analysis of judgment of partners' emotions during e-mail and face-to-face communication. *Journal of Science Education in Japan, 29*(5), 354–365.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*(1), 32–38.

Kiesler, S., Siegel, J., & McGuire, T. W. (1984). Social psychological aspects of computer-mediated communication. *The American Psychologist*. Retrieved from http://psycnet.apa.org/record/1985-27678-001?casa_token=qx7q_6sto90AAAAA:Kugom-T2X-rVGijd9o1apc8n3CzOLO6VilvN7NxSksk52-t_ukX4IQiwh8oxoruyG3ZZ6s0Vv0Uz-4uBmdGWBds

Kingsbury, D. (1968). *Manipulating the amount of information obtained from a person giving directions*. Harvard University.

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron, 61*(1), 140–151.

Klucharev, V., Smidts, A., & Fernández, G. (2008). Brain mechanisms of persuasion: How "expert power" modulates memory and attitudes. *Social Cognitive and Affective Neuroscience, 3*(4), 353–366.

Krach, S., Blümel, I., Marjoram, D., Lataster, T., Krabbendam, L., Weber, J., … Kircher, T. (2009). Are women better mindreaders? Sex differences in neural correlates of mentalizing detected with functional MRI. *BMC Neuroscience, 10*, 9.

Krauss, R. M., & Fussell, S. R. (1991). Perspective-taking in communication: Representations of others' knowledge in reference. *Social Cognition, 9*(1), 2–24.

Kuhlen, A. K., Bogler, C., Brennan, S. E., & Haynes, J.-D. (2017). Brains in dialogue: Decoding neural preparation of speaking to a conversational partner. *Social Cognitive and Affective Neuroscience, 12*(6), 871–880.

Kumkale, G. T., Albarracin, D., & Seignourel, P. J. (2010). The effects of source credibility in the presence or absence of prior attitudes: Implications for the design of persuasive communication campaigns. *Journal of Applied Social Psychology, 40*(6), 1325–1356.

Lee, C.-C., Black, M., Katsamanis, A., Lammert, A. C., Baucom, B. R., Christensen, A., ... Narayanan, S. S. (2010). Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In *Eleventh Annual Conference of the International Speech Communication Association*. isca-speech.org. Retrieved from http://www.isca-speech.org/archive/interspeech_2010/i10_0793.html

Lombardo, M. V., Barnes, J. L., Wheelwright, S. J., & Baron-Cohen, S. (2007). Self-referential cognition and empathy in autism. *PLoS One, 2*(9), e883.

Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Baron-Cohen, S., & MRC AIMS Consortium. (2011). Specialization of right temporo-parietal junction for mentalizing and its relation to social impairments in autism. *NeuroImage, 56*(3), 1832–1838.

Ma, Ning, Marie Vandekerckhove, Kris Baetens, Frank Van Overwalle, Ruth Seurinck, and Wim Fias. (2012). Inconsistencies in Spontaneous and Intentional Trait Inferences. *Social Cognitive and Affective Neuroscience 7*(8), 937–50.

Masten, C. L., Morelli, S. A., & Eisenberger, N. I. (2011). An fMRI investigation of empathy for "social pain" and subsequent prosocial behavior. *NeuroImage, 55*(1), 381–388.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences of the United States of America, 98*(20), 11832–11835.

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience, 8*(6), 623–631.

Mitchell, J. P. (2009). Inferences about mental states. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 364*(1521), 1309–1316.

Mønster, D., Håkonsson, D. D., Eskildsen, J. K., & Wallot, S. (2016). Physiological evidence of interpersonal dynamics in a cooperative production task. *Physiology & Behavior, 156*, 24–34.

Newman-Norlund, S. E., Noordzij, M. L., Newman-Norlund, R. D., Volman, I. A. C., de Ruiter, J. P., Hagoort, P., & Toni, I. (2009). Recipient design in tacit communication. *Cognition, 111*(1), 46–54.

Nimchinsky, E. A., Gilissen, E., Allman, J. M., Perl, D. P., Erwin, J. M., & Hof, P. R. (1999). A neuronal morphologic type unique to humans and great apes. *Proceedings of the National Academy of Sciences of the United States of America, 96*(9), 5268–5273.

Noordzij, M. L., Newman-Norlund, S. E., de Ruiter, J. P., Hagoort, P., Levinson, S. C., & Toni, I. (2009). Brain mechanisms underlying human communication. *Frontiers in Human Neuroscience, 3*, 14.

Norman, E., Tjomsland, H. E., & Huegel, D. (2016). The distance between us: Using construal level theory to understand interpersonal distance in a digital age. *Frontiers in Digital Humanities, 3*, 5.

O'Donnell, M. B., Bayer, J. B., Cascio, C. N., & Falk, E. B. (2017). Neural bases of recommendations differ according to social network structure. *Social Cognitive and Affective Neuroscience, 12*(1), 61–69.

O'Donnell, M. B., Falk, E. B., & Lieberman, M. D. (2015). Social in, social out: How the brain responds to social language with more social language. *Communication Monographs, 82*(1), 31–63.

Paal, T., & Bereczkei, T. (2007). Adult theory of mind, cooperation, Machiavellianism: The effect of mindreading on social relations. *Personality and Individual Differences, 43*(3), 541–551.

Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2018). Similar neural responses predict friendship. *Nature Communications, 9*(1), 332.

Petty, R. E., Cacioppo, J. T., & Heesacker, M. (1984). Central and peripheral routes to persuasion: Application to counseling. *Social Perception in Clinical and Counseling Psychology, 2*, 59–60.

Ridinger, G., & McBride, M. (2017). *Theory-of-mind ability and cooperation*. Retrieved from http://economics.ucr.edu/seminars_colloquia/2017-18/economic_theory/McBride%20paper%20for%201%2031%2018%20seminar.pdf

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *NeuroImage, 22*(4), 1694–1703.

Rubin, D. L., & Rafoth, B. A. (1986). Social cognitive ability as a predictor of the quality of expository and persuasive writing among college freshmen. *Research in the Teaching of English, 20*, 9–21.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In J. Schenkein (Ed.), *Studies in the organization of conversational interaction* (pp. 7–55). New York, NY: Academic Press.

Sally, D., & Hill, E. (2006). The development of interpersonal strategy: Autism, theory-of-mind, cooperation and fairness. *Journal of Economic Psychology, 27*(1), 73–97.

Sanfey, A. G. (2007). Social decision-making: Insights from game theory and neuroscience. *Science, 318*(5850), 598–602.

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology, 55*, 87–124.

Schiller, D., Freeman, J. B., Mitchell, J. P., Uleman, J. S., & Phelps, E. A. (2009). A neural mechanism of first impressions. *Nature Neuroscience, 12*(4), 508–514.

Schmälzle, R., Häcker, F. E. K., Honey, C. J., & Hasson, U. (2015). Engaged listeners: Shared neural processing of powerful political speeches. *Social Cognitive and Affective Neuroscience, 10*(8), 1137–1143.

Scholz, C., Baek, E. C., O'Donnell, M. B., Kim, H. S., Cappella, J. N., & Falk, E. B. (2017). A neural model of valuation and information virality. *Proceedings of the National Academy of Sciences of the United States of America, 114*(11), 2881–2886.

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences of the United States of America, 111*(43), E4687–E4696.

Silvia, P. J. (2005). Deflecting reactance: The role of similarity in increasing compliance and reducing resistance. *Basic and Applied Social Psychology, 27*(3), 277–284.

Sproull, L., & Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communication. *Management Science, 32*(11), 1492–1512.

Spunt, R. P., & Lieberman, M. D. (2012). An integrative model of the neural systems supporting the comprehension of observed emotional behavior. *NeuroImage, 59*(3), 3050–3059.

Stallen, M., Smidts, A., & Sanfey, A. G. (2013). Peer influence: Neural mechanisms underlying in-group conformity. *Frontiers in Human Neuroscience, 7*, 50.

Stephens, G. J., Silbert, L. J., & Hasson, U. (2010). Speaker–listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences, 107*(32), 14425–14430.

Sternthal, B., Phillips, L. W., & Dholakia, R. (1978). The persuasive effect of scarce credibility: A situational analysis. *Public Opinion Quarterly, 42*(3), 285–314.

Stolk, A., D'Imperio, D., di Pellegrino, G., & Toni, I. (2015). Altered communicative decisions following ventromedial prefrontal lesions. *Current Biology: CB, 25*(11), 1469–1474.

Taijfel, H. (1970). Experiments in intergroup discrimination. *Scientific American, 223*(5), 96–102.

Tang, H., Mai, X., Wang, S., Zhu, C., Krueger, F., & Liu, C. (2016). Interpersonal brain synchronization in the right temporo-parietal junction during face-to-face economic exchange. *Social Cognitive and Affective Neuroscience, 11*(1), 23–32.

Traxler, M. J., & Gernsbacher, M. A. (1993). Improving written communication through perspective-taking. *Language and Cognitive Processes, 8*(3), 311–334.

Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2008). The neural substrates of in-group bias: A functional magnetic resonance imaging investigation. *Psychological Science, 19*(11), 1131–1139.

Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2011). Modulation of the fusiform face area following minimal exposure to motivationally relevant faces: Evidence of in-group enhancement (not out-group disregard). *Journal of Cognitive Neuroscience, 23*(11), 3343–3354.

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30*(3), 829–858.

Vanlangendonck, F., Willems, R. M., & Hagoort, P. (2018). Taking common ground into account: Specifying the role of the mentalizing network in communicative language production. *PLoS One, 13*(10), e0202943.

Vettel, J. M., Lauharatanahirun, N., Wasylyshyn, N., Roy, H., Fernandez, R., Cooper, N., … Garcia, J. O. (2019). Translating driving research from simulation to interstate driving with realistic traffic and passenger interactions. In *Advances in human factors in simulation and modeling* (pp. 126–138). New York, NY: Springer International Publishing.

Welborn, B. L., Lieberman, M. D., Goldenberg, D., Fuligni, A. J., Galván, A., & Telzer, E. H. (2016). Neural mechanisms of social influence in adolescence. *Social Cognitive and Affective Neuroscience, 11*(1), 100–109.

Wilson, E. J., & Sherrell, D. L. (1993). Source effects in communication and persuasion research: A meta-analysis of effect size. *Journal of the Academy of Marketing Science, 21*(2), 101.

Woodside, A. G., & Davenport, J. W., Jr. (1974). The effect of salesman similarity and expertise on consumer purchasing behavior. *Journal of Marketing Research, 11*(2), 198–202.

Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., & Hasson, U. (2017). How we transmit memories to other brains: Constructing shared neural representations via communication. *Cerebral Cortex, 27*(10), 4988–5000.

Zerubavel, N., Bearman, P. S., Weber, J., & Ochsner, K. N. (2015). Neural mechanisms tracking popularity in real-world social networks. *Proceedings of the National Academy of Sciences of the United States of America, 112*(49), 15072–15077.

Zhang, Z., Ye, Q., Law, R., & Li, Y. (2010). The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. *International Journal of Hospitality Management, 29*(4), 694–700.

Part VI
Mentalizing in Self-Referential Processing and Emotion

Tangled Representations of Self and Others in the Medial Prefrontal Cortex

Robert S. Chavez

It comes as a surprise to no one to learn that humans are a fundamentally social species. We have evolved a set of mental capacities that enable us to interact with and navigate through our deeply complex sociality. As such, our minds are equipped with a powerful capacity to understand the thoughts, intentions, and mental states of other people. This process is called *mentalizing*, and it is a defining characteristic of our species. We infer thoughts, feelings, and beliefs in the minds of other people, often without intention and seemingly automatically. Indeed, these processes are so effortless that we spontaneously engage them even when perceiving inanimate objects behaving as if they have minds (Heider & Simmel, 1944). Undoubtedly, the capacity for mentalizing and the degree to which we think about others is a central characteristic of our species (Baumeister & Leary, 1995), and much of our mental lives are spent mentalizing or preparing to process information about social agents in our environment (Lieberman, 2013; Meyer, Davachi, Ochsner, & Lieberman, 2018).

Despite the fact that we spend so much time thinking of other people, there is another target of our mind's eye that we may think about even more: Ourselves. Indeed, it could be argued that the human capacity for a rich and deep sense of self is equally as defining to our species and central to our mental lives as mentalizing about others. The relationship between the way we think of ourselves and the way we think of others surrounds several central questions in social psychology and neuroscience. What is the link between self and person knowledge? Are there common cognitive processes and brain systems that underlie both mentalizing and self-reference? Where do they diverge and what does that mean for each of them? Here, I will briefly review some of the evidence showing that there is both overlap and divergence in the neural systems supporting each of these processes and will outline some approaches for potentially disentangling some of these issues.

R. S. Chavez (✉)
Department of Psychology, University of Oregon, Eugene, OR, USA
e-mail: rchavez@uoregon.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_31

Do Mentalizing and Self-Representation Share Common Neural Systems?

Theories about the nature of the self and its relationship to others have been a part of psychology since its inception. William James (1890) defined many of the concepts that would come to be central to the study of self. James distinguished material and social aspects of the self—such as a person's body parts and their outward persona toward others—from the internal aspects of the self that must be introspected and evaluated in order to define one's values and personality. James' ideas held the self as the centerpiece of the theorizing others more directly related self-knowledge with mentalizing. Cooley's (1902) concept of the so-called *looking glass self* postulated that the principle way by which we gain self-knowledge is gained through reflection and feedback we receive about ourselves from others in our social word—"Everyone tells me I work very hard, so I must be a hard worker." These early ideas about the relationship between the self and the way we think about other people would ultimately come to be among the most popular topics in social cognition and eventually social neuroscience.

Since the early 2000s, researchers have identified a system of brain regions that are involved in mentalizing, theory of mind, and person perception. These regions largely fall within the default mode network and include the temporal parietal junction, posterior cingulate cortex, and temporal pole areas. Although there is debate about the degree to which there is a *primary* area involved in mentalizing or theory of mind per se (Mitchell, 2007; Saxe & Powell, 2006), perhaps the most consistently implicated region in mentalizing and related social cognitive processes is the medial prefrontal cortex (MPFC). Indeed, studies have shown that the dorsal MPFC is implicated in tasks involving theory of mind (Mitchell, 2007), perceptions of animacy (Wheatley, Milleville, & Martin, 2007), and general person knowledge and social cognitive processes (Mitchell, Heatherton, & Macrae, 2002) using classic univariate fMRI methods. More recently, multivariate and data-driven methods have implicated areas within the MPFC for processing familiar faces (Visconti di Oleggio Castello, Halchenko, Guntupalli, Gors, & Gobbini, 2017), responding preferentially to naturalistic social interaction (Wagner, Kelley, Haxby, & Heatherton, 2016), and encoding identity-specific trait information (Hassabis et al., 2014). Indeed, these findings point to a critical role for the MPFC in mentalizing and social cognition.

Similarly, self-referential processing has also been most consistently linked to activation in the MPFC. Initial studies on the neural basis of self-representation compared activation of making trait-judgements for the self, relative to a familiar but unknown other (e.g., political figures), and found that a portion of the ventral MPFC was most consistently activated during these tasks (Kelley et al., 2002; Macrae, Moran, Heatherton, Banfield, & Kelley, 2004). Since then, researchers found that the MPFC also responds more to thinking about the self than to thinking about a close friend (Heatherton et al., 2006) and is recruited when using information about the self to make decisions about the preferences of another person (Tamir & Mitchell, 2010). Importantly, MPFC responds to both implicit and explicit self-relevant information (Moran, Heatherton, & Kelley, 2009), suggesting that the recruitment of the MPFC for self-representation is not simply driven by overt task demands.

Taken together, there is clear evidence that both mentalizing and self-referential processes recruit portions of the MPFC. Indeed, researchers have posited that the overlap in cortical regions serving these seemingly disparate phenomena is evidence of a shared psychological process underlying each of them (Mitchell, 2009). However, there remains issues with this strict interpretation.

The Dorsal/Ventral Gradient

Despite the fact that both mentalizing and self-representation share some common cortical real estate, their respective spatial distributions in the MPFC are not completely congruent. Specifically, many researchers posited that mentalizing more consistently activates dorsal portions of the MPFC, whereas self-reference appears to activate more ventral regions of the MPFC. This observation has led many to believe that regions recruited for processing information about the self differ from regions implicated in processing information about others. Indeed, this account was fueled by the results of two influential meta-analyses (Denny, Kober, Wager, & Ochsner, 2012; Wagner, Haxby, & Heatherton, 2012), which both largely support the conclusion that self and other processes can be dissociated as you move dorsally along the cortico-frontal midline.

The idea that a cortical gradient neatly separates self from others is compelling, but there are several issues with this interpretation. First, multiple studies have shown that the ventral MPFC is also responsive to other people, particularly when we know them personally (Krienen, Tu, & Buckner, 2010) or they are similar to ourselves (Mitchell, Macrae, & Banaji, 2006). Similarly, dorsal MPFC has been shown to be engaged during self-reference when considering the certainty or epistemic information of self-judgements. This is in contrast to the ventral MPFC which is more tuned to the importance or evaluative content of those judgements (D’Argembeau et al., 2012). These and similar findings cast doubt on the idea that a strict dorsal/ventral gradient is the most appropriate way to account for the differences in self/other cognition.

Also clouding the strict self/other gradient distinction is an often overlooked but important detail in both the Denny et al. (2012) and Wagner et al. (2012) meta-analyses. In each of these papers, the authors did not include results for regions that were ventral to a $Z = -10$ in Montreal Neurological Institute (MNI) standard space. As such, neither of these meta-analyses considered the more ventral portions of the MPFC and medial orbitofrontal cortex, which we know process information about self (Hughes & Beer, 2013), empathy (Morelli, Rameson, & Lieberman, 2014), and person perception in affective contexts (Chavez & Heatherton, 2015b). Moreover, subregions of the ventral MPFC and OFC are areas frequently implicated in reward processes, economic valuation, and other information that is not social cognitive per

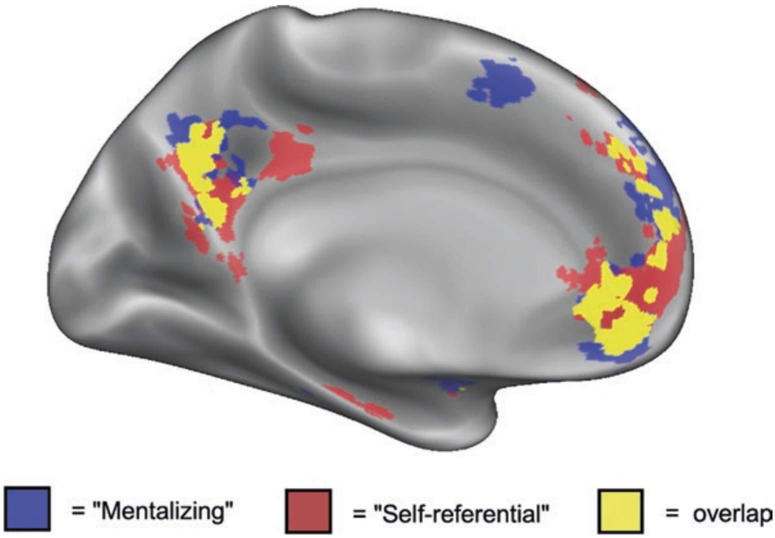


Fig. 1 Overlap of mentalizing and self-referential processing is largest in the ventral medial prefrontal cortex based on text-based meta-analysis in Neurosynth (Yarkoni et al., 2011). Red areas represent results from 166 studies using the term "self-referential," blue areas represent results from 115 studies using the term "mentalizing," and the yellow areas represent the overlap between the two. Each map was generated using the uniformity test procedure to highlight areas that showed a high probability of being activated across studies given the term

se (Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008). Indeed, these issues are even further compounded by the lack of a standardized nomenclature for labeling regions and subregions within the MPFC, leading different researchers to call the same region a different name and vice versa.

Nonetheless, overall there does seem to be some degree of separability among the brain regions involved in mentalizing and self-reference. At the same time, there remains a considerable amount of cortical overlap in the brain regions, particularly with the MPFC, that contribute to both self-reference and general social cognition (see: Fig. 1). If this is indeed the case, how can we begin to understand how this brain region is working to contribute to each of these domains?

Mixed Selectivity

Even in its simplest form, when engaging in mentalizing, we are required to take in and distill a bevy of complex information. We need to perceive and identify the presence of another agent in the environment, infer the actions or intentions of that agent, measure those behaviors against our knowledge of that person and their current context, and finally perform a calculation to integrate this information together and allow us to act accordingly. Even if you only consider part of the mentalizing network, how might it be possible that such complex information processing could be distilled into a region such as the MPFC, especially when that region is involved in processing other diverse types of information?

Unlike the relative homogeneity of brain regions involved in lower-level sensory processes, the prefrontal cortex is thought to integrate diverse forms of information. Theories of general prefrontal cortex function highlight its role in serving goal pursuit, information flow, and top-down control (Miller & Cohen, 2001). One of the key insights from these theories is how the prefrontal cortex orchestrates these processes by considering both intrinsically and extrinsically generated signals, such as those involved in thinking of others and ourselves. Thus, a critical part of the function of the prefrontal cortex during both mentalizing and self-representation may be to integrate incoming information from another person with endogenous signals that are already present or generated simultaneously. One idea of how neurons in the prefrontal cortex accomplish this is through exhibiting so-called *mixed selectivity*.

Typically, neurons exhibit response functions that are tuned to selectively respond to one particular type of stimulus. Classic examples of this include line orientation response tuning in primary visual cortex and specific pitch in primary auditory cortex. However, relative to other parts of the brain, the prefrontal cortex is disproportionately comprised of mixed selectivity neurons—single cells that respond to a mixture of multiple task-relevant features (Rigotti et al., 2013). Mixed selectivity neurons combine to contribute high-level cognitive representations through both linear and nonlinear response tunings to multiple stimulus aspects (Rigotti et al., 2013). Thus, even at the level of the individual cells, many prefrontal cortical neurons are not category-specific.

Given the heterogeneous responses even within single cells of prefrontal neurons, it remains a possibility that this complex information tuning can be observed at higher levels too. Though it can be difficult to extrapolate principles from individual neurons to information coded at the region-level, it is very unlikely that portions of the MPFC are selective for mentalizing, self-representation, or most other complex constructs. This hints at the promise that we can use tools like fMRI to better understand how constructs are dissociated in the brain, even when they are represented in overlapping areas. Indeed, relevant information related to these processes may be coded in broader-scale neural ensemble patterns that can be detected at the voxel level or across multivoxel response patterns. If this is indeed the case, we need to consider multivariate information that is embedded both within local activity within the MPFC as well as how the MPFC coordinates information across distributed neural systems involved in mentalizing, self-representation, and the more basic psychological processes on which each are built.

Leveraging Multivariate Methods

One of the major developments in cognitive neuroscience in the past decade has been the shift away from univariate brain mapping studies and toward multivariate pattern analysis (MVPA) methods, such as pattern decoding, representational similarity analysis (RSA), and voxelwise encoding methods (Haxby, Connolly, &

Guntupalli, 2014). Indeed, in many areas of cognitive neuroscience, these methods are now standard practice and are common throughout the literature. Though slower to pick up speed in social neuroscience, there has been a recent surge of studies employing MVPA methods to study social cognition as well (for review see: Wagner, Chavez, & Broom, 2019). Like many issues in social cognition, these efforts began by studying more basic social informational processes in sensory systems.

One of the most important sources of information about others is, of course, the face. The ventral visual stream, including the fusiform face area, has long been established as the core set of regions responsible for processing information about faces (Haxby, Hoffman, & Gobbini, 2000; Kanwisher, McDermott, & Chun, 1997). However, a harder challenge for researchers studying faces has been *identity decoding*—successfully predicting which specific person a subject is seeing based on neural activity patterns. However, using a combination of decoding, RSA, and hyperalignment, a study by Visconti di Oleggio Castello et al. (2017), showed that identity decoding was indeed possible for both familiar and unfamiliar faces. Importantly, however, these effects were not confined to the ventral visual stream, and identity-specific information was able to be decoded from the dorsal MPFC, even for unfamiliar faces. It is also worth noting that in this experiment participants were engaged in a simple visual oddball-detection task and not in a mentalizing task, per se. Given these findings, this may reflect an automatic propensity to engage mentalizing systems for the purpose of gathering information about others. Alternatively, it may simply reflect the general role of the dorsal MPFC in integrating identity related information about conspecifics. Additional research will be necessary to tease apart each of these possibilities.

Face perception is important for gathering information about others, but not sufficient for mentalizing. Instead, we also need to consider the thoughts and traits of others. Can this information be detected in MPFC with MVPA? To test this, Hassabis et al. (2014) had subjects learn personalities of four different identities and think of each of them in different situational contexts. They showed that a searchlight MVPA decoder could differentiate each identity in the dorsal MPFC. Importantly, the personality traits assigned to each identity were counterbalanced across subjects, suggesting that individuals are indeed coding for elements of the traits, rather than other low-level visual features of the identities. Similarly, Thornton and Mitchell (2017) asked subjects to think of 20 personally familiar people in a variety of situations. Using an RSA approach, they found that a similarity structure based on identity could be found throughout the social brain and default network, including dorsal and ventral MPFC. Both the Hassabis et al. (2014) and Thornton and Mitchell (2017) studies show that multivariate patterns in the MPFC can be used to understand how we think about the minds of other people.

Aspects of self-representation can also be gleaned from multivariate response patterns in the MPFC. In one study, my colleagues and I trained an MVPA classifier to dissociate positive and negative affect using visual images (Chavez, Heatherton, & Wagner, 2017). We then applied the classification boundary from this procedure to brain activation patterns when subjects were thinking about themselves or thinking of their best friend. We found that this cross-domain decoding approach could successfully dissociate thinking about the self from thinking about a friend in the ventral MPFC. Similarly, a study by Yankouskaya et al. (2017) found a similar result, except in this study they train a classifier to dissociate high versus low reward value. Like the studies on social cognition, these studies indicate that information about the self is reflected in multivariate response patterns in the MPFC and that it may be possible to predict these patterns based on their underlying affective components.

However, because these MVPA studies on self-representation used other people as the contrasting condition, they also begin to more directly address the question, "What dimension separates self and others in the MPFC?" These studies suggest a compelling candidate answer to this question: positive affect or reward-related processing. Clearly, the representation between self and others in the MPFC cannot be accounted for entirely by a dissociation in valence or reward. However, these and similar multivariate methods may provide a way of testing additional possible cognitive dimensions that can further disentangle the representations of self and others, even within overlapping portions of the MPFC.

Fusing Multiple Modalities

MVPA methods may provide a useful means to ask how information is represented within a local region. However, we also know that long-range information coordination is critical for supporting higher-order psychological processes, including mentalizing and self-referential processing. These processes are thought to be supported by both the local and long-range connectivity between systems via the brain's white matter pathways. Indeed, researchers have found that humans have disproportionately greater prefrontal white matter relative to other primates (Schoenemann, Sheehan, & Glotzer, 2005). This hints at the possibility that incorporating information about structural connectivity may inform our understanding of the functional role of the prefrontal cortex, including the MPFC, in mentalizing and self-reference.

In humans, structural connectivity is typically assessed in vivo using diffusion magnetic resonance imaging (dMRI) using tractography methods. There are now several studies that have used dMRI to understand various aspects of social cognition (Wang & Olson, 2018), including phenomena related to mentalizing, such as empathy (Parkinson & Wheatley, 2012). However, social neuroscientists are often not interested in white matter per se, but rather its functional relevance. This has motivated some groups to take a more comprehensive approach, utilizing both dMRI and fMRI within the same study. For example, one study had subjects complete a standard trait judgement task and measured the task-based functional connectivity between the MPFC and the ventral striatum while making people made positive evaluations of themselves using fMRI. In the same subjects, they also measured the white matter connectivity of the same regions using dMRI. They found that individual differences in short-term state self-esteem were related to frontostriatal connectivity using fMRI, whereas long-term trait self-esteem was more strongly related to frontostriatal white matter integrity using dMRI (Chavez & Heatherton, 2015a). Although this approach of using each modality separately can yield interesting insights into mentalizing and self-reference, the biggest promise of using multimodal methods is when they can be systematically combined to provide a greater understanding of the functional specialization of an area based on its underlying structural characteristics.

Using a highly innovative paradigm, Saygin et al. (2012) sought to test the question of whether the structural connectivity of a region could predict its functional specificity; in this case, face selective cortex in the fusiform gyrus. To do this, these researchers scanned individuals doing a standard facial perception task using fMRI before acquiring a high-quality dMRI scan for calculating structural connectivity with tractography. Next, using a leave-one-subject-out cross-validation procedure, they trained a model to predict face selective voxels in the fusiform based only on the structural connectivity of these voxels to the rest of the brain. They found that structural connectivity measures could accurately predict the location of the fusiform face in each subject's brain, and that these predictions even outperformed the group-level average for capturing each subject's idiosyncratic face selective region in each subject's own brain. Thus, the results from Saygin et al. (2012) demonstrate that, indeed, the structural connectivity of a system may constrain and predict the functional specialization of that system. Moreover, they also demonstrated the utility of systematically combining fMRI and dMRI modalities to inform the processes whereby structure begets function.

This approach may provide a roadmap for how to conduct a similar investigation into how and under what conditions regions of the MPFC are specialized for mentalizing or self-reference. To date, there has not been a study attempting to do this. However, there are additional issues that will make this an even bigger challenge. The main issue is that, unlike the ventral visual stream, the MPFC does not have highly selective patches of cortex that are dedicated to mentalizing, self-reference, or most other processes. Because the estimates of structural connectivity of a region is fixed within the individual, it provides the same information for predicting binary boundaries of masks where they overlap. Put differently, if two entirely congruent regions are both marked as mentalizing and self-reference areas, predicting their specialization using brain structure will not be possible using the approach analogues to the fusiform face area localization mask in Saygin et al. (2012). Nonetheless, these challenges are not insurmountable and remain ripe for future investigations using carefully crafted tasks to elicit both self-reference and mentalizing processes.

This is just one example of the large possibility space we are afforded when combining dMRI with fMRI to understand how the MPFC supports and untangles mentalizing from other phenomena. There are dozens of other neuroimaging and psychophysiological modalities that could be combined together in a systematic way to better understand these issues too. And though there are technical and practical challenges to employing multiple methods, systematically combining information across modalities will help us achieve a more comprehensive understanding of what is being computed within the MPFC and which other brain systems are acting with it in concert during mentalizing as compared to other psychological processes.

Shifting Paradigms

Advanced analytic and acquisition methods like the ones described above are certainly going to be a critical part of further understanding how mentalizing is represented in the MPFC. However, there is no amount of technical sophistication that can substitute for informative experimental designs. Indeed, in addition to the recent advancements in methodological approaches, there has also been an increasing interest in employing innovative paradigms to shed new light on how social information is processed in the brain. Two popular approaches in this vein are the use of naturalistic stimuli and social network analysis.

A persistent issue in much of social neuroscience is that the MRI machine is just about the most unnatural context one can think of to study social cognition. Subjects are lying on their backs inside of a plastic donut as it shrieks at them while they try to half-focus on some artificial, humdrum task. Although there is no way to completely circumvent these issues, it is possible to get a bit of a boost in ecological validity through the use of naturalistic stimuli such as movies and audio narratives to understand how the brain is processing different elements of social cognition. For example, Wagner et al. (2016) showed subjects clips from the Hollywood film *Matchstick Men* while in the scanner. Using a data-driven reverse-correlation procedure pioneered by Hasson (2004), they found that activity in the dorsal MPFC during natural viewing was preferentially engaged during scenes with multiple characters interacting on the screen. These results underscore the importance of the dorsal MPFC in processing information about mental state inference, especially during social interactions.

Other studies have used fMRI to understand how the brain encodes broader social context information using social network analysis. In a pioneering study by Parkinson, Kleinbaum, and Wheatley (2017), an entire incoming class of business school students were recruited to estimate the friendship and familiarity properties of every other subject in the network. Next, a subset of these subjects was brought into the scanner to view short video clips of each person in the network introducing themselves. Brain activity during these clips was then used to predict whether subjects spontaneously encoded social network information. They found that, indeed, several regions could dissociate the processing of social distance between subjects as well as measures of target subject's social network metrics of brokerage. Relevant to the current discussion of mentalizing, they found that *eigenvector centrality* —a measure of prestige based on how well-connected one is to other well-connected people—was spontaneously encoded throughout the social brain network, including the MPFC. These results provide some exciting insights, suggesting that the MPFC codes not only for information about the social agents right in front of our eyes, but is also embedding that information into our representation of broader social contexts.

Although studies using naturalistic stimuli and social network analyses are beginning to offer new insight into the nature of social cognition in the MPFC, they do not directly address the issue of overlapping representations of mentalizing with self-representation in the MPFC. To address some of these issues, my collaborators and I have proposed using round-robin designs—a design in which each study participant is both a perceiver and target for every other subject in the study—to directly relate how much of our sense of self is reflected in the brains of others within our close-knit social groups. In a forthcoming study, we used this round-robin approach in a standard self/other trait judgement paradigm to demonstrate that brain activity in the MPFC during self-reflection in a target subject could be predicted from brain activity in the same area in the brains of others when they are thinking about that target (Chavez & Wagner, under review). In a different study, Zerubavel, Bearman, Weber, and Ochsner (2015) used a round-robin design to present photos of each participant within their social network to one another and found that activity in canonical social perception, including the MPFC, tracked sociometric popularity and regions associated with valuation tracked target popularity.

Together, these studies suggest that information about the self is being reflected in the brains of others in our social groups, which may help us to better understand how person knowledge shifts from information-gathering when we encounter strangers, to a sense of confidence in knowing another person as we become familiar with them. Moreover, future work could also combine round-robin paradigms with the naturalistic stimuli and network analysis approaches described above. This would help to even further understand the similarities and differences of self/other processing and how each of these psychological phenomena are being represented in the MPFC.

Conclusion

An abundance of studies has made it clear that the MPFC is a critical region supporting our ability to consider the thoughts and motivations of others that are required for mentalizing. However, it is also clear that overlapping portions of the MPFC are also involved in processing information about the self and that the MPFC is not specialized solely for the purposes of processing of information about either mentalizing or self-representation. Indeed, given this region's functional heterogeneity, even at the level of individual neurons, it is going to require even greater ingenuity to further tease apart the underlying similarities and differences in the ways that the MPFC computes information about others and ourselves. This chapter has highlighted some of the ways in which we are beginning to take these steps and aimed to provide fodder for even more thorough investigation of this topic.

Finally, throughout this chapter, I have been discussing mentalizing and self-reference as if they are simple, discrete psychological categories unto themselves. I hope it is obvious, however, that each of the processes is built on a host of more basic psychological processes—What is mentalizing without attention? What is a sense of identity without memory?—Several of these more basic processes undoubtedly are playing out in concert in order to orchestrate the types of behavior and cognition that get labeled as "mentalizing." To the degree that we want to understand what the MPFC is actually computing during mentalizing, we are going to have to understand its cognitive and affective component parts and what exactly they are computing. This may, in turn, help us distinguish the brain basis of mentalizing from processes such as self-reflection and other closely related phenomena. The tools to accomplish these goals are right in front of us, but they will require a shift toward more complex experimental paradigms wedded with increased methodological sophistication, similar to the paradigms discussed in this chapter. We have come a long way in identifying the brain areas involved in social cognition, including the MPFC. But the journey to deeply understanding the biological basis of mentalizing is only just beginning.

References

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*(3), 497–529. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7777651

Chavez, R. S., & Heatherton, T. F. (2015a). Multimodal frontostriatal connectivity underlies individual differences in self-esteem. *Social Cognitive and Affective Neuroscience, 10*(3), 364–370. https://doi.org/10.1093/scan/nsu063

Chavez, R. S., & Heatherton, T. F. (2015b). Representational similarity of social and valence information in the medial pFC. *Journal of Cognitive Neuroscience, 27*(1), 73–82. https://doi.org/10.1162/jocn_a_00697

Chavez, R. S., Heatherton, T. F., & Wagner, D. D. (2017). Neural population decoding reveals the intrinsic positivity of the self. *Cerebral Cortex, 27*(11), 5222–5229. https://doi.org/10.1093/cercor/bhw302

Cooley, C. H. (1902). Looking-glass self. In *The production of reality: Essays and readings on social interaction* (Vol. 6). Thousand Oaks, CA: Sage.

D'Argembeau, A., Jedidi, H., Balteau, E., Bahri, M., Phillips, C., & Salmon, E. (2012). Valuing one's self: Medial prefrontal involvement in epistemic and emotive investments in self-views. *Cerebral Cortex, 22*(3), 659–667.

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience, 24*(8), 1742–1752. https://doi.org/10.1162/jocn_a_00233

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience, 28*(22), 5623–5630.

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2014). Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex, 24*(8), 1979–1987. https://doi.org/10.1093/cercor/bht042

Hasson, U. (2004). Intersubject synchronization of cortical activity during natural vision. *Science, 303*(5664), 1634–1640. https://doi.org/10.1126/science.1089506

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience, 37*(1), 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*(6), 223–233. https://doi.org/10.1016/S1364-6613(00)01482-0

Heatherton, T. F., Wyland, C. L., Macrae, C. N., Demos, K. E., Denny, B. T., & Kelley, W. M. (2006). Medial prefrontal activity differentiates self from close others. *Social Cognitive and Affective Neuroscience, 1*(1), 18–25. https://doi.org/10.1093/scan/nsl001

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology, 57*(2), 243. https://doi.org/10.2307/1416950

Hughes, B. L., & Beer, J. S. (2013). Protecting the self: The effect of social-evaluative threat on neural representations of self. *Journal of Cognitive Neuroscience, 25*(4), 613–622.

James, W. (1890). *The principles of psychology* (Vol. 1). New York, NY: Holt.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience, 17*(11), 4302–4311. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9151747

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience, 14*(5), 785–794. https://doi.org/10.1162/08989290260138672

Krienen, F. M., Tu, P.-C., & Buckner, R. L. (2010). Clan mentality: Evidence that the medial prefrontal cortex responds to close others. *Journal of Neuroscience, 30*(41), 13906–13915. https://doi.org/10.1523/JNEUROSCI.2180-10.2010

Lieberman, M. D. (2013). *Social: Why our brains are wired to connect*. Oxford, England: OUP.

Macrae, C. N., Moran, J. M., Heatherton, T. F., Banfield, J. F., & Kelley, W. M. (2004). Medial prefrontal activity predicts memory for self. *Cerebral Cortex, 14*(6), 647–654. https://doi.org/10.1093/cercor/bhh025

Meyer, M. L., Davachi, L., Ochsner, K. N., & Lieberman, M. D. (2018). Evidence that default network connectivity during rest consolidates social information. *Cerebral Cortex, 29*(5), 1910–1920. https://doi.org/10.1093/cercor/bhy071

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*, 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Mitchell, J. P. (2007). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex, 18*(2), 262–271.

Mitchell, J. P. (2009). Social psychology as a natural kind. *Trends in Cognitive Sciences, 13*(6), 246–251. https://doi.org/10.1016/j.tics.2009.03.008

Mitchell, J. P., Heatherton, T. F., & Macrae, C. N. (2002). Distinct neural systems subserve person and object knowledge. *Proceedings of the National Academy of Sciences of the United States of America, 99*(23), 15238–15243. https://doi.org/10.1073/pnas.232395699

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50*(4), 655–663. https://doi.org/10.1016/j.neuron.2006.03.040

Moran, J. M., Heatherton, T. F., & Kelley, W. M. (2009). Modulation of cortical midline structures by implicit and explicit self-relevance evaluation. *Social Neuroscience, 4*(3), 197–211. https://doi.org/10.1080/17470910802250519

Morelli, S. A., Rameson, L. T., & Lieberman, M. D. (2014). The neural components of empathy: Predicting daily prosocial behavior. *Social Cognitive and Affective Neuroscience, 9*(1), 39–47. https://doi.org/10.1093/scan/nss088

Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2017). Spontaneous neural encoding of social network position. *Nature Human Behaviour, 1*(5), 72. https://doi.org/10.1038/s41562-017-0072

Parkinson, C., & Wheatley, T. (2012). Relating anatomical and social connectivity: White matter microstructure predicts emotional empathy. *Cerebral Cortex, 24*, 614. https://doi.org/10.1093/cercor/bhs347

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature, 497*(7451), 585–590. https://doi.org/10.1038/nature12160

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science, 17*(8), 692–699.

Saygin, Z. M., Osher, D. E., Koldewyn, K., Reynolds, G., Gabrieli, J. D. E., & Saxe, R. R. (2012). Anatomical connectivity patterns predict face selectivity in the fusiform gyrus. *Nature Neuroscience, 15*(2), 321–327.

Schoenemann, P. T., Sheehan, M. J., & Glotzer, L. D. (2005). Prefrontal white matter volume is disproportionately larger in humans than in other primates. *Nature Neuroscience, 8*(2), 242–252. https://doi.org/10.1038/nn1394

Tamir, D. I., & Mitchell, J. P. (2010). Neural correlates of anchoring-and-adjustment during mentalizing. *Proceedings of the National Academy of Sciences of the United States of America, 107*(24), 10827–10832. https://doi.org/10.1073/pnas.1003242107

Thornton, M. A., & Mitchell, J. P. (2017). Consistent neural activity patterns represent personally familiar people. *Journal of Cognitive Neuroscience, 29*(9), 1583–1594. https://doi.org/10.1162/jocn_a_01151

Visconti di Oleggio Castello, M., Halchenko, Y. O., Guntupalli, J. S., Gors, J. D., & Gobbini, M. I. (2017). The neural representation of personally familiar and unfamiliar faces in the distributed system for face perception. *Scientific Reports, 7*(1), 12237. https://doi.org/10.1038/s41598-017-12559-1

Wagner, D. D., Chavez, R. S., & Broom, T. W. (2019). Decoding the neural representation of self and person knowledge with multivariate pattern analysis and data-driven approaches. *Wiley Interdisciplinary Reviews: Cognitive Science, 10*(1), e1482. https://doi.org/10.1002/wcs.1482

Wagner, D. D., Haxby, J. V., & Heatherton, T. F. (2012). The representation of self and person knowledge in the medial prefrontal cortex. *Wiley Interdisciplinary Reviews: Cognitive Science, 3*(4), 451–470. https://doi.org/10.1002/wcs.1183

Wagner, D. D., Kelley, W. M., Haxby, J. V., & Heatherton, T. F. (2016). The dorsal medial prefrontal cortex responds preferentially to social interactions during natural viewing. *The Journal of Neuroscience, 36*(26), 6917–6925. https://doi.org/10.1523/JNEUROSCI.4220-15.2016

Wang, Y., & Olson, I. R. (2018). The original social network: White matter and social cognition. *Trends in Cognitive Sciences, 22*(6), 504–516. https://doi.org/10.1016/j.tics.2018.03.005

Wheatley, T., Milleville, S. C., & Martin, A. (2007). Understanding animate agents. *Psychological Science, 18*(6), 469–474. https://doi.org/10.1111/j.1467-9280.2007.01923.x

Yankouskaya, A., Humphreys, G., Stolte, M., Stokes, M., Moradi, Z., & Sui, J. (2017). An anterior–posterior axis within the ventromedial prefrontal cortex separates self and reward. *Social Cognitive and Affective Neuroscience, 12*(12), 1859–1868. https://doi.org/10.1093/scan/nsx112

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., Wager, T. D., & Van Essen, D. C. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods, 8*(8), 665–670. https://doi.org/10.1038/NMETH.1635

Zerubavel, N., Bearman, P. S., Weber, J., & Ochsner, K. N. (2015). Neural mechanisms tracking popularity in real-world social networks. *Proceedings of the National Academy of Sciences, 112*(49), 15072–15077.

Why Don't You Like Me? The Role of the Mentalizing Network in Social Rejection

Razia S. Sahi and Naomi I. Eisenberger

Rejection hurts. Although this phrase is typically meant metaphorically, a body of evidence suggests that social rejection may hurt literally, much like physical pain (for reviews, see Eisenberger, 2012; Eisenberger & Lieberman, 2005). But not every instance of social rejection cuts deeply. When we find out that the romantic interest who's been acting distant recently lost a loved one, or that a mutual friend who's been giving us the cold shoulder is anxious around people they don't know well, this knowledge can alter our interpretations of their behavior and lead us to feel less hurt by their actions. It seems that being socially rejected hurts not just because someone ignores or dismisses us, but rather because we feel that their rejection has something to do with how they think and feel about us. We notice how someone is acting—distant, cold, uninterested—and we wonder why they might be acting this way: "why don't they like me?" The pain of rejection, or at least some forms of rejection, seems to be inherently tied to the way that we interpret another person's thoughts or feelings.

It makes sense that the pain of social rejection relies to some extent on how we perceive the intentions of the person rejecting us. Indeed, even the experience of physical pain is more intense when we perceive our pain to be intentionally caused by someone else (Wegner & Gray, 2008). Despite this intuitive connection between the experience of social rejection and the process of thinking about and trying to understand someone else's thoughts and feelings, a process referred to as "mentalizing" (Frith & Frith, 2006), very little neuroscience research has explicitly examined the role of mentalizing in the experience of social rejection.

In this review, we explore evidence from the current literature to examine the possible role of mentalizing in the experience of social rejection. To do this, we first turn to meta-analyses investigating the neural bases of social rejection to examine whether parts of the mentalizing network are also active during the experience of

R. S. Sahi · N. I. Eisenberger (✉)
Department of Psychology, University of California, Los Angeles, Los Angeles, CA, USA
e-mail: rsahi1@ucla.edu; neisenbe@ucla.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_32

rejection (Cacioppo et al., 2013; Vijayakumar, Cheng, & Pfeifer, 2017). Next, we assess whether developmental changes in mentalizing, such as those during early childhood and adolescence, are associated with changes in sensitivity to rejection (e.g., Rochat, 2003; Somerville, 2013). Then, we examine whether individuals who demonstrate compromised mentalizing, such as those with schizophrenia or autism, exhibit reduced sensitivity to rejection (e.g., Bauminger & Kasari, 2000; Gradin, Waiter, Kumar, Stickle, & Milders, 2012). Finally, we summarize some future directions building on the possibility of an inherent link between mentalizing and the experience of social rejection. We suggest that the available evidence supports a potential role of the mentalizing network in feeling the pain of social rejection, such that understanding another person's mental state may be what allows us to understand and process rejection.

Are Mentalizing Regions Active During the Experience of Social Rejection?

Research suggests that we have such a strong aversion to social rejection that even rejection by a stranger, from whom we have little to gain or lose, can cause us significant distress. For example, Eisenberger et al. (2003) published the first study to use a paradigm called "Cyberball" to induce feelings of social rejection in participants who were laying alone in a functional magnetic resonance imaging (fMRI) scanner. Cyberball is a virtual ball-toss game that involves three avatars passing a ball back and forth. The participant believes that one avatar represents themselves, while the other two avatars represent the other players in the game. However, in reality, there are no other players; instead, the program is designed to include or exclude the participant from the ball-tossing game. Initially, the participant is included in the game; however, in an exclusion condition, the participant's avatar no longer receives the ball from the other avatars. Even in this context where participants are not physically around other people and have little to lose by being excluded, participants report feeling distressed by the rejection.

The Cyberball paradigm has consistently elicited feelings of social distress in participants across populations (e.g., Gradin et al., 2012; Groschwitz, Plener, Groen, Bonenberger, & Abler, 2016; Masten et al., 2011), across modified versions of the paradigm (e.g., DeWall et al., 2012.; Onoda et al., 2009), and even in studies where participants know they are not really playing the game with other people (Zadro, Williams, & Richardson, 2004). The pain of rejection is so salient that researchers have theorized that the experience of rejection may have piggybacked on the physical pain system, borrowing the pain signal to denote the potential for broken social bonds, warning us to avoid them (Eisenberger & Lieberman, 2004; MacDonald & Leary, 2005; Panksepp, 2004). By allowing us to detect the threat of exclusion, which can restrict access to resources, social support, and other protective factors, social pain can help us adaptively navigate the social world and maintain the relationships that promote our well-being (Eisenberger & Lieberman, 2005).

Since the first study using Cyberball, countless studies have examined the neural mechanisms underlying feelings of rejection. While many of these studies have used the Cyberball paradigm, others have used more personally relevant paradigms, such as thinking about rejection by recounting a recent romantic break-up (Fisher, Brown, Aron, Strong, & Mashek, 2010; Kross, Berman, Mischel, Smith, & Wager, 2011). These studies have predominantly focused on examining the neural regions associated with the affective (unpleasant-feeling) component of physical pain (i.e., dorsal anterior cingulate cortex (dACC), anterior insula (AI)), and have not directly investigated the relationship between mentalizing and social rejection. However, by examining meta-analyses of such studies, we can investigate whether there is evidence of a consistent role for mentalizing-related neural regions in the experience of social rejection. First, we will briefly describe the neural network typically associated with mentalizing. Then, we will assess the extent to which this network seems to play a role in experiencing social rejection.

The Mentalizing Network

Social cognition researchers draw a distinction between the ways that we understand *how* others do things and *why* they do things. Understanding how someone does something involves grasping the mechanisms of an action, whereas understanding why someone does something involves reasoning about their mental states, i.e., mentalizing (Spunt, Falk, & Lieberman, 2010). In the case of social rejection, understanding *how* someone is passing a ball back and forth to another person is experientially distinct from understanding *why* they are passing the ball to the other person. While *how-thinking* doesn't seem to play a role in our own feelings about the game or the other players, *why-thinking* can lead us to wonder why we are not receiving the ball from others. Such thoughts can lead to hurt feelings, self-doubt, offense, embarrassment, and a host of other negative emotions.

While mentalizing can sometimes lead to negative emotional experiences, it allows us to understand the intentions, goals, and emotions of those around us, which informs how we behave and communicate with others, and facilitates our ability to collaborate with others towards achieving joint goals (Saxe, 2006). Given the clear distinction between *how-thinking* and *why-thinking*, and the importance of reasoning about mental states in our everyday experience, extensive research has documented the neural bases of this social cognitive process. The "mentalizing network," as it has come to be called, is typically thought to include regions such as the temporoparietal junction (TPJ), dorsomedial prefrontal cortex (dmPFC), precuneus, posterior superior temporal sulcus (pSTS), and the temporal poles (Frith & Frith, 2006; Gallagher & Frith, 2003; Lieberman, 2010), with some evidence suggesting that ventromedial prefrontal cortex (vmPFC) contributes to related social cognitive processes (Lieberman, Straccia, Meyer, Du, & Tan, 2019). Each of these neural regions is thought to assist with processing different sorts of information that collectively facilitate understanding the mental states of others. While the contributing role of each region is still not clearly understood, mentalizing is thought to consist of various sub-processes such as interpreting human motion in terms of goals or intentions, representing mental states, and shared-understanding of others' emotional states (Saxe, 2006).

Meta-analyses of Social Rejection

While no research has directly examined the link between the mentalizing network and the experience of social rejection, there have been several meta-analyses on the neural bases of social rejection that can indicate whether the mentalizing network tends to be active during the experience of social rejection. For example, Cacioppo et al. (2013) conducted a meta-analysis of neuroimaging studies, including 12 Cyberball studies using 244 participants, and 3 studies that involved thinking about a recent unwanted break-up from a romantic partner using 60 participants. During the latter romantic rejection studies, participants were exposed to photographs of their ex-partners and were asked to relive the memory of the unwanted break-up (e.g., Fisher et al., 2010; Kross et al., 2011). This meta-analysis found that during Cyberball, but not during the break-up task, there was significant activity in dmPFC, a neural region that has consistently been shown to play a role in mentalizing (Lieberman et al., 2019; Saxe, 2006). Notably, the analysis of Cyberball studies included a much larger sample ($N = 244$) than that of the romantic rejection studies ($N = 60$), so it is possible that the meta-analysis of romantic rejection studies was relatively underpowered to detect significant activity within neural regions such as dmPFC. Furthermore, participants in the romantic rejection studies may already have reflected on and come to understand their past rejection before taking part in the break-up task, whereas those playing Cyberball may have been trying to understand why they were being rejected during the task itself, thus recruiting greater mentalizing resources.

A more recent meta-analysis conducted by Vijayakumar et al. (2017) extended Cacioppo et al.'s work to examine 40 studies, including 1122 participants who underwent different types of social exclusion tasks, including the social judgment and chatroom tasks. In both of these tasks, participants evaluate unfamiliar peers based on their photographs, and then receive feedback about how they were evaluated by those peers.

This meta-analysis also specifically examined 857 participants from Cyberball tasks to compare the patterns of activation from this task to other social exclusion tasks more generally. They found that across all social exclusion studies, there was significant neural activation in regions such as the precuneus, dmPFC, and vmPFC. Meanwhile, Cyberball specifically tended to elicit activation in precuneus and vmPFC. Although not definitive, as the role of mentalizing in social rejection was not specifically examined in these studies, these results are consistent with a role of the mentalizing network in the experience of social rejection across a variety of exclusion tasks.

Based on the findings reviewed above, the mentalizing network seems to play a role in the experience of social rejection. In the subsequent two sections, we will expand our investigation to consider whether individual differences in the mentalizing network are associated with variations in sensitivity to social rejection. First, we will consider whether developmental changes in mentalizing are associated with changes in sensitivity to social rejection. If children first become sensitive to social rejection (i.e., self-conscious) when they develop the ability to mentalize, this association would suggest a link between mentalizing and feeling the pain of social rejection. Furthermore, if adolescents who demonstrate particularly high sensitivity to social rejection also demonstrate enhanced mentalizing, this association would also suggest a potential role of mentalizing in the experience of social rejection. Then, we will review research examining whether impairments in the ability to mentalize, such as those observed in schizophrenia and autism, are associated with differences in sensitivity to social rejection. Finally, we will summarize some future directions relating to the role of the mentalizing network in feelings of social rejection.

Are Developmental Changes in Mentalizing Associated with Changes in Sensitivity to Rejection?

The way that we think about ourselves in relation to other people undergoes significant changes from early childhood through adolescence and adulthood, resulting in changes in emotional responsivity to social events across development. Two notable developmental changes in emotional responsivity linked to a growing concern for one's social relationships are (a) the emergence of self-conscious emotions and sensitivity to social rejection in early childhood (i.e., around 3–8 years old) (Rochat, 2003), and (b) heightened negative emotional responsivity to social rejection during adolescence (i.e., the time between puberty and adulthood) (Somerville, 2013). Interestingly, both of these developmental milestones are marked by significant changes in the mentalizing network. In what follows, we will describe the potential link between mentalizing and sensitivity to social rejection in terms of these two developmental time periods.

The Emergence of Mentalizing in Early Childhood

Around the age of 4–6, children become increasingly adept at understanding the thoughts and feelings of other people, even when they conflict with something they know about the external world (i.e., false beliefs) (Frith & Frith, 2003). To illustrate, if an object is placed in one location in front of a third party, but moved to a second location without that person's knowledge, we would infer that the person believes that the object is in the original location since they did not witness the object being moved. However, children before the age of 4 typically fail to recognize the person's false belief that the object is in the original location, instead reporting that the person must know that the object has been moved. When children develop the ability to perform this complex mentalizing task, it demonstrates their ability to infer that other people have their own thoughts and feelings that are separate from one's own thoughts and feelings (Rochat, 2003).

In an effort to better understand the development of such mentalizing abilities in early childhood, some research has examined the neural correlates of performing false belief tasks in this age group. This research finds that children who can perform these tasks exhibit increased neural responsivity in regions associated with mentalizing, including dmPFC and TPJ, as compared to children who have not yet developed this ability (Liu, Sabbagh, Gehring, & Wellman, 2009; Sabbagh, Bowman, Evraire, & Ito, 2009). These findings suggest that the maturation of the mentalizing network plays a role in the emergence of complex mentalizing abilities during this time period.

Interestingly, the emergence of mentalizing abilities in early childhood overlaps with the emergence of self-consciousness. Self-consciousness includes the experience of social emotions such as embarrassment and shame, and is attributed to children's growing awareness that other people hold some perception of them. In other words, as children begin to think about how others think and feel about them, they also begin to experience negative feelings about being perceived undesirably (Frith & Frith, 2003; Rochat, 2003). The development of these self-conscious thoughts and feelings has been explained in terms of an evolutionary need to affiliate with others and the resultant fear of social rejection that supports our ability to maintain social bonds (Rochat, 2009). More specifically, once children learn to understand the thoughts and feelings of others, they also recognize the potential for negative social evaluation that could lead to social rejection, resulting in negative social emotions that tend to modulate social behavior, such as embarrassment.

This developmental association between the ability to understand the thoughts and feelings of others and exhibiting a fear of social rejection through self-conscious behavior indicates a potential link between mentalizing and experiencing the pain of social rejection. Indeed, this research suggests that in the absence of complex mentalizing ability, children may not understand and process social rejection, and thus may not experience the same levels of emotional distress as healthy adults when rejected. However, further research is necessary to explicitly explore this association between mentalizing and social rejection in early childhood.

Hyper-mentalizing in Adolescence

A defining feature of adolescence is the importance of peer and romantic relationships. The importance of these social relationships is thought to increase adolescents' social sensitivity such that social information becomes particularly salient

(Somerville, 2013). Because of fluctuations in social relationships during this time, social rejection is common (Wang, Iannotti, & Nansel, 2009). Thus, adolescents are not only more likely to experience social rejection, but are also more likely to demonstrate heightened negativity in response to the experience of rejection.

Researchers have investigated adolescents’ emotional responsivity to social rejection in a number of ways, including Cyberball, social judgment, and chatroom tasks (Silk et al., 2012; Somerville, Heatherton, & Kelley, 2006; Williams, Cheung, & Choi, 2000). Compared to adults, adolescents tend to report worse mood and anxiety following rejection (Sebastian, Viding, Williams, & Blakemore, 2010), expect less favorable positive feedback from their peers (Moor, van Leijenhorst, Rombouts, Crone, & van der Molen, 2010), and show greater pupillary dilation in response to rejection (Silk et al., 2012). Given this sensitivity to social rejection, information about the thoughts and feelings of others, particularly as this information relates to the self and one’s social relationships, would be especially important to adolescents. To this end, we might expect the mentalizing network to be more responsive to social information in adolescents than adults.

Indeed, research suggests that adolescents recruit mPFC to a greater extent than adults during Cyberball (Sebastian et al., 2011), and during tasks that involve considering the thoughts and intentions of others (Burnett, Sebastian, Cohen Kadosh, & Blakemore, 2011). Researchers have also found greater functional connectivity in adolescents than in adults between regions of the mentalizing network, including pSTS and TPJ, and anterior rostral mPFC during tasks that involve thinking about social emotions (Burnett & Blakemore, 2009). Since the brain continues to mature throughout development, researchers have suggested that heightened sensitivity to rejection during this age range may be due to the continuing maturation of the mPFC during this time (Blakemore, 2008). Alternatively, adolescents may recruit mPFC to a greater extent because of the importance of social information at this age.

If the mentalizing network plays a role in understanding and processing social rejection, then greater sensitivity to social rejection may be associated with increased activity in the mentalizing network, potentially explaining why certain individuals are more sensitive to social rejection in the first place. The research described in this section suggests that heightened emotional sensitivity to rejection during adolescence could be related to heightened activity in the mentalizing network. Taken together with research suggesting that self-consciousness and the fear of social rejection first emerge when children develop complex mentalizing abilities, this developmental literature supports a potential role of the mentalizing network in processing and experiencing social rejection.

Thus far, we have reviewed whether neural regions associated with mentalizing are also active during the experience of social rejection, and whether developmental changes in sensitivity to rejection relate to neural activity in the mentalizing network. In the next section, we consider whether impairments in the ability to mentalize may be associated with changes in sensitivity to social rejection. If the mentalizing network plays a role in understanding and processing social rejection, then impairments in the ability to mentalize may be associated with decreased sensitivity to social rejection.

Are Impairments in Mentalizing Associated with Reduced Sensitivity to Rejection?

Two clinical disorders that are characterized by significant impairments in the ability to infer emotional and mental states are schizophrenia (Brüne, 2005; Frith & Corcoran, 1996) and autism (American Psychiatric Association, 1994). In this section, we consider how deficits in mentalizing may be affecting the way that individuals with schizophrenia and autism process social rejection. If mentalizing is important for feeling socially rejected, then there could be evidence for decreased sensitivity to social rejection in these two populations. While neuroscience research explicitly testing this relationship in clinical populations is limited, the available research suggests that deficits in the mentalizing network may be contributing to abnormalities in how individuals with schizophrenia and autism respond to social rejection, as well as consequential difficulties in social interaction stemming from an inability to properly process social evaluative cues.

Social Rejection in Schizophrenia

Schizophrenia is often accompanied by symptoms such as delusions and hallucinations involving social content, and deficits in motivation and social skills, ultimately leading to difficulty in social interaction that impedes everyday functioning (MacDonald & Leary, 2005). These social challenges are often explained by mentalizing deficits in this population in so far as a failure to understand the thoughts and feelings of others can lead individuals with schizophrenia to perceive threat in the absence of harmful intentions. Furthermore, failure to understand others' mental states generally makes it difficult for individuals with schizophrenia to regulate their social behavior and interactions in accordance with social feedback.

In an attempt to better understand such deficits, a growing body of research has investigated abnormalities in the structure and function of the mentalizing network in patients with schizophrenia (Benedetti et al., 2009; Mier et al., 2010; Park et al., 2011). However, few studies have investigated this network in patients explicitly during social rejection. One such study sheds some light on how abnormalities in the mentalizing network might shape the recognition and processing of social rejection. This study used a version of the Cyberball task in which exclusion was parametrically modulated (participants receive the ball some percentage of the time), as opposed to being dichotomous (participant either receives the ball proportional to other players in the game or does not receive the ball at all) (Gradin et al., 2012).

In response to social exclusion, the control group demonstrated increased activation in the vmPFC, a region sometimes implicated in mentalizing, and the ventral anterior cingulate cortex (vACC), a region that has been shown to activate to social exclusion and to be associated with social distress (Bolling et al., 2011b; Gunther Moor et al., 2012). Meanwhile, patients with schizophrenia failed to modulate activity in these regions in accordance with percentage of exclusion, with greater positive symptom severity corresponding to lower modulation of activity. Within the schizophrenic group, but not the control group, stronger responses to social exclusion in the vmPFC were associated with greater self-reported social distress.

These findings suggest abnormal neural responsivity to social exclusion in the mentalizing network in schizophrenia. First, the schizophrenic group's failure to modulate activity within the vACC may point to a lack of sensitivity to social rejection. Moreover, although the schizophrenic group failed to modulate activity in the vmPFC overall, greater activity in the vmPFC was associated greater self-reported distress from social exclusion, suggesting that those with better mentalizing ability may have felt more social pain. Thus, individuals with schizophrenia exhibited abnormalities within the mentalizing network during social rejection, as well as diminished activation in regions of the brain associated with social distress during rejection. Impairments in the ability to accurately interpret the thoughts and feelings of others may hinder this population's ability to detect rejection when it is truly occurring, thereby inhibiting appropriate modulation of social distress in response to rejection cues, resulting in reduced sensitivity to true rejection.

A second fMRI study examining social rejection in a schizophrenic population used a virtual reality handshake task to induce feelings of social rejection in this population (Lee et al., 2014). In this task, participants' physical hand movements controlled an avatar on the screen such that when participants raised their hand, their avatar would offer a handshake to another avatar on the screen. Depending on the condition, the stranger avatar either exhibited friendly body language and accepted the handshake (i.e., acceptance), or unfriendly body language and refused the handshake (i.e., rejection).

The results of this study indicated abnormalities within the mentalizing network in the schizophrenic group during social rejection, providing some insight into the neural mechanisms underlying the social deficits associated with this disorder. First, as compared to the control group, the schizophrenia group exhibited significantly lower activity during rejection versus acceptance in pSTS, a region within the mentalizing network associated with identifying the motivations behind bodily movements (Saxe, 2006), with greater symptom severity corresponding to less activity in pSTS. This finding suggests that individuals with schizophrenia may not be able to properly recruit the neural regions necessary to accurately interpret social cues that provide information about the goals and intentions of other people. Second, the schizophrenia group exhibited significantly greater activity during rejection versus acceptance in left vmPFC, suggesting that individuals with schizophrenia may be recruiting certain social cognitive processes to a greater extent than healthy individuals during social rejection. While these two results initially seem conflicting, together they indicate abnormal neural responsivity to social rejection in schizophrenia in regions associated with processing social information, suggesting a potential role of mentalizing deficits in how individuals with schizophrenia experience social rejection.

In terms of differences in self-reported feelings of rejection, this study found that the schizophrenia group reported greater feelings of rejection during acceptance than the control group, but exhibited no difference in such feelings during rejection. While this finding does not inform whether individuals with schizophrenia experienced differential levels of distress in response to rejection, it does help explain positive psychotic symptoms in schizophrenia such as delusions about persecution in the absence of real threat (Park et al., 2011).

In sum, the failure to appropriately modulate regions associated with mentalizing, as well as those associated with social distress, during social exclusion provides some explanation for positive psychotic symptoms in schizophrenia. In failing to accurately recognize and interpret social cues, individuals with schizophrenia may demonstrate blunted affect in cases of actual rejection, and demonstrate heightened affect in the absence of social threat. While no research has explicitly tested how impairments in mentalizing potentially impact sensitivity to social rejection in schizophrenia, the studies described in this section suggest that abnormalities in the mentalizing network may be associated with atypical responsivity to social rejection in schizophrenia.

Social Rejection in Autism

A core feature of autism is impairment in social interaction, which leads to difficulty in forming and maintaining social relationships (American Psychiatric Association, 1994; Baron-Cohen, 2010). When examined objectively, for example through social network analysis applied to children within classrooms, individuals with autism tend to face more instances of social rejection. For example, they tend to experience lower centrality, less acceptance, less companionship, and less reciprocity in the social networks they inhabit (Chamberlain, Kasari, & Rotheram-Fuller, 2007).

Despite this difference in both quality and quantity of social relationships, some research suggests that children with autism do not tend to report greater feelings of loneliness or related sadness than their peers (Bauminger et al., 2008; Chamberlain et al., 2007). This finding has been interpreted in terms of the lack of awareness that individuals with autism may have about their experience of rejection. For example, Chamberlain et al. (2007) proposed that mentalizing deficits in autism might leave children unable to recognize the shortcomings of their social relationships. In line with this hypothesis, parents of children with autism tend to report that their children seemed generally oblivious about social cues that would signal social inclusion or exclusion (Chamberlain et al., 2007).

One study that sheds some light on the experience of social rejection in autism examined loneliness and friendship in a population of high-functioning children with autism (Bauminger & Kasari, 2000). They found that while children with autism desired friendship like typical children, they experienced greater loneliness as captured by the loneliness rating scale (e.g., “I have nobody to talk to in class”). Upon inspecting how children with autism and typical children defined loneliness, both groups defined loneliness in terms of being alone (i.e., having no one to play with), but children with autism were much less likely than typical children to define loneliness in terms of negative emotional feelings such as sadness, depression, or fear. These results suggest that children with autism recognized that they were left out of friendships or activities, but did not necessarily internalize this rejection in a way that affected their emotional states. In other words, while children with autism did not want to be alone and recognized when they were alone, they did not seem to experience the pain of rejection in the same way as typical children who described *feelings* of loneliness.

A later study designed to explicitly explore how adolescents with autism experience social rejection as compared to healthy adolescents found that while both groups experienced similar levels of distress and anxiety during Cyberball, only the healthy adolescents showed significantly lower self-reported mood after rejection as compared with baseline and inclusion conditions (Sebastian, Blakemore, & Charman, 2009). In other words, there seemed to be a lasting effect of social rejection in healthy individuals, but no reduction in later mood in adolescents with autism. This finding suggests that while individuals with autism may recognize and respond to social rejection in the moment, they may not process and internalize rejection in the same way as typical individuals, resulting in a lower likelihood of *feelings* of loneliness over time.

While there is no neuroscience research explicitly testing whether mentalizing impairments in autism are associated with diminished sensitivity to rejection in this population, multiple neuroimaging studies have examined how social rejection is experienced in autism (Bolling et al., 2011a; 2011b; Masten et al., 2011; McPartland et al., 2011). Across these studies, there were no differences in immediate self-reported responses to rejection between the autism group and the control group. However, individuals with autism demonstrated lower neural responsivity to rejection as compared with the control groups, particularly in the vACC and right AI, regions often associated with experiencing social distress (Masten et al., 2011; McPartland et al., 2011). These results suggest, again, that individuals with autism recognize and dislike social rejection, but that they may not have the same negative feelings associated with being rejected as do typically developing individuals.

A recent meta-analysis of the functional neural correlates of social and non-social tasks in autism similarly demonstrates differential neural responsivity to social stimuli more broadly in this population. This meta-analysis examined 24 studies of social processes (e.g., theory of mind, face perception) and 15 studies of non-social processes (e.g., attention control, working memory) in adults with autism (Di Martino et al., 2009). This analysis revealed decreased likelihood of activation in anterior rostral mPFC, a region implicated in self-referential processing (Lieberman et al., 2019), as well as regions associated with social distress, such as the dACC and right AI, during social tasks versus non-social tasks in individuals with autism as compared to typical individuals. Since this meta-analysis did not include studies particularly related to social rejection, further research is necessary to examine the role of the mentalizing network in sensitivity to social rejection in autism. However, research in this area is consistent with diminished sensitivity to social rejection in autism, both in terms of lasting feelings about rejection, and in terms of immediate neural responsivity to social rejection. Of course, it is not known whether a reduced ability to mentalize precipitates diminished rejection sensitivity or whether a heightened sensitivity to rejection leads to a compensatory reduction in mentalizing; this would need to be examined in future studies. In the next and final section, we will offer some conclusory remarks and potential future directions building on the possibility of an inherent link between mentalizing and the experience of social rejection.

Conclusion and Future Directions

While limited research has examined the connection between the mentalizing network and social rejection, we have summarized a body of evidence suggesting that the mentalizing network plays a potential role in how we understand and process social rejection. First, meta-analyses investigating the neural underpinnings of social rejection with a focus on pain-related regions such as the dACC and AI have found consistent activation of mentalizing regions, such as regions within the mPFC and precuneus, in the experience of social rejection across a variety of rejection paradigms. Second, developmental research suggests that self-consciousness and fear of social rejection emerge in early childhood when children first develop the ability to perform complex mentalizing tasks, which is marked by increased activation in mentalizing regions. Additionally, adolescents who tend to exhibit heightened sensitivity to social rejection also tend to exhibit heightened activation within the mentalizing network in response to social rejection. Third, clinical populations that are characterized by deficits in the ability to mentalize, including schizophrenia and autism, tend to demonstrate decreased sensitivity to social rejection in the form of inappropriately modulated affect in schizophrenia, blunted affect following rejection in autism, and abnormal patterns of activity during rejection in neural regions associated with social distress, such as the vACC, in both schizophrenia and autism.

A breadth of research suggests a possible inherent link between mentalizing and the experience of social rejection. However, further research explicitly testing the association between this neural network and social-emotional experiences is necessary in order to explain whether mentalizing is required for understanding and processing social rejection, and to explain the mechanism by which mentalizing potentially affects emotional experience. In addition to building on the clinical and developmental research we have summarized in this chapter, there are several additional avenues for research that could illuminate the role of mentalizing in social rejection. For example, research examining functional connectivity between mentalizing regions during social tasks can investigate whether these regions are more connected during social rejection. One such study suggests that connectivity between regions within the mentalizing network, including dmPFC, vmPFC, precuneus, and TPJ, increases during social exclusion compared to social inclusion (Schmälzle et al., 2017). Further research is necessary to examine the consistency of this result, but this preliminary finding provides a promising direction for future research in this area.

Another potentially fruitful area for future research involves examining how individuals' mentalizing activity during social rejection may change as a function of their vulnerability to social rejection. Individuals who are at greater risk of rejection, or have more to lose if they are socially rejected, might devote greater resources to mentalizing about others so that they can better predict and thus avoid possible experiences of rejection in the future. For example, individuals who are low in social status are more vulnerable to rejection, since lower social status can mean less access to resources, and therefore greater risk of being excluded, as well as greater cost of exclusion. Such individuals seem to recruit mentalizing resources to a greater extent than typical or high status individuals during social tasks (Muscatell et al., 2012). As a second example, individuals with less dense friendship networks, suggesting less relationship stability and social support provisions (Lin, 2002), have shown greater functional connectivity within the mentalizing network (greater coupling between left and right TPJ) during social exclusion (Schmälzle et al., 2017). Such preliminary findings suggest a potential link between vulnerability to social rejection and mentalizing that could be an interesting an avenue for future research.

Ultimately, we have suggested that mentalizing may play a role in understanding and processing social rejection insofar as understanding how someone else thinks and feels about you may underpin the pain of feeling rejected. While on the one hand, understanding someone else's thoughts and feelings may allow you to interpret their behavior as lacking malice (e.g., They are just in a bad mood today.), it can also provide you with insight into how others view you (e.g., They don't like me.) Wondering why someone doesn't like us may bring us to the undesirable conclusion that there is something about us that is disagreeable to others. However, understanding and processing this rejection seems to be part of learning to build and maintain social bonds—without which we would suffer significantly greater pains than the pain of a single rejection.

References

American Psychiatric Association. Task Force on DSM-IV. (1994). Dsm-iv sourcebook (Vol. 1). American Psychiatric Pub.

Baron-Cohen, S. (2010). The empathizing-systematizing (E-S) theory of autism: A cognitive developmental account. In *The Wiley-Blackwell handbook of childhood cognitive development* (2nd ed., pp. 626–639). Oxford, England: Wiley-Blackwell. https://doi.org/10.1002/9781444325485.ch24

Bauminger, N., & Kasari, C. (2000). Loneliness and friendship in high-functioning children with autism. *Child Development, 71*(2), 447–456. https://doi.org/10.1111/1467-8624.00156

Bauminger, N., Solomon, M., Aviezer, A., Heung, K., Gazit, L., Brown, J., & Rogers, S. J. (2008). Children with autism and their friends: A multidimensional study of friendship in high-functioning autism spectrum disorder. *Journal of Abnormal Child Psychology, 36*(2), 135–150. https://doi.org/10.1007/s10802-007-9156-x

Benedetti, F., Bernasconi, A., Bosia, M., Cavallaro, R., Dallaspezia, S., Falini, A., … Smeraldi, E. (2009). Functional and structural brain correlates of theory of mind and empathy deficits in schizophrenia. *Schizophrenia Research, 114*(1), 154–160. Retrieved from https://www.sciencedirect.com/science/article/pii/S0920996409002916

Blakemore, S. J. (2008). The social brain in action. *Nature Reviews Neuroscience, 9*(4), 267. Retrieved from https://www.nature.com/articles/nrn2353

Bolling, D. Z., Pitskel, N. B., Deen, B., Crowley, M. J., McPartland, J. C., Kaiser, M. D., … Pelphrey, K. A. (2011a). Enhanced neural responses to rule violation in children with autism: A comparison to social exclusion. *Developmental Cognitive Neuroscience, 1*(3), 280–294. https://doi.org/10.1016/j.dcn.2011.02.002

Bolling, D. Z., Pitskel, N. B., Deen, B., Crowley, M. J., McPartland, J. C., Mayes, L. C., & Pelphrey, K. A. (2011b). Dissociable brain mechanisms for processing social exclusion and rule violation. *NeuroImage, 54*(3), 2462–2471. https://doi.org/10.1016/j.neuroimage.2010.10.049

Brüne, M. (2005). "Theory of mind" in schizophrenia: A review of the literature. *Schizophrenia Bulletin, 31*(1), 21–42. https://doi.org/10.1093/schbul/sbi002

Burnett, S., & Blakemore, S. J. (2009). Functional connectivity during a social emotion task in adolescents and in adults. *European Journal of Neuroscience, 29*(6), 1294–1301. https://doi.org/10.1111/j.1460-9568.2009.06674.x

Burnett, S., Sebastian, C., Cohen Kadosh, K., & Blakemore, S. J. (2011). The social brain in adolescence: Evidence from functional magnetic resonance imaging and behavioural studies. *Neuroscience and Biobehavioral Reviews, 35*, 1654. https://doi.org/10.1016/j.neubiorev.2010.10.011

Cacioppo, S., Frum, C., Asp, E., Weiss, R. M., Lewis, J. W., & Cacioppo, J. T. (2013). A quantitative meta-analysis of functional imaging studies of social rejection. *Scientific Reports, 3*, 2027. https://doi.org/10.1038/srep02027

Chamberlain, B., Kasari, C., & Rotheram-Fuller, E. (2007). Involvement or isolation? The social networks of children with autism in regular classrooms. *Journal of Autism and Developmental Disorders, 37*(2), 230–242. https://doi.org/10.1007/s10803-006-0164-4

Dewall, C. N., Masten, C. L., Powell, C., Combs, D., Schurtz, D. R., & Eisenberger, N. I. (2012). Do neural responses to rejection depend on attachment style? An fMRI study. *Social Cognitive and Affective Neuroscience, 7*(2), 184–192. https://doi.org/10.1093/scan/nsq107

Di Martino, A., Ross, K., Uddin, L. Q., Sklar, A. B., Castellanos, F. X., & Milham, M. P. (2009). Functional brain correlates of social and nonsocial processes in autism spectrum disorders: An activation likelihood estimation meta-analysis. *Biological Psychiatry, 65*(1), 63–74. https://doi.org/10.1016/j.biopsych.2008.09.022

Eisenberger, N. I. (2012). The pain of social disconnection: Examining the shared neural underpinnings of physical and social pain. *Nature Reviews Neuroscience, 13*(6), 421–434. https://doi.org/10.1038/nrn3231

Eisenberger, N. I., & Lieberman, M. D. (2004). Why rejection hurts: A common neural alarm system for physical and social pain. *Trends in Cognitive Sciences, 8*(7), 294. https://doi.org/10.1016/j.tics.2004.05.010

Eisenberger, N. I., & Lieberman, M. D. (2005). Why it hurts to be left out the neurocognitive overlap between physical and social pain. In J. Forgas, K. D. Williams, & W. von Hippel (Eds.), *The social outcast: Ostracism, social exclusion, rejection, and bullying* (p. 130). East Sussex, UK: Psychology Press.

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. Science, 302(5643), 290-292.

Fisher, H. E., Brown, L. L., Aron, A., Strong, G., & Mashek, D. (2010). Reward, addiction, and emotion regulation systems associated with rejection in love. *Journal of Neurophysiology, 104*, 51–60. https://doi.org/10.1152/jn.00784.2009

Frith, C. D., & Corcoran, R. (1996). Exploring "theory of mind" in people with schizophrenia. *Psychological Medicine, 26*(3), 521. https://doi.org/10.1017/S0033291700035601

Frith, C. D., & Frith, U. (2006). Minireview the neural basis of mentalizing. *Neuron, 50*, 531–534. https://doi.org/10.1016/j.neuron.2006.05.001

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society B: Biological Sciences, 358*, 459. https://doi.org/10.1098/rstb.2002.1218

Gallagher, H. L., & Frith, C. D. (2003). Network and psychological effects in URBAN MOVEMENT.pdf, *7*(2), 77–83. https://doi.org/10.1016/S1364-6613(02)00025-6

Gradin, V. B., Waiter, G., Kumar, P., Stickle, C., & Milders, M. (2012). Abnormal neural responses to social exclusion in schizophrenia. *PLoS One, 7*(8), 42608. https://doi.org/10.1371/journal.pone.0042608

Groschwitz, R. C., Plener, P. L., Groen, G., Bonenberger, M., & Abler, B. (2016). Differential neural processing of social exclusion in adolescents with non-suicidal self-injury: An fMRI study. *Psychiatry Research - Neuroimaging, 255*, 43–49. https://doi.org/10.1016/j.pscychresns.2016.08.001

Gunther Moor, B., Güroğlu, B., Op de Macks, Z. A., Rombouts, S. A. R. B., Van der Molen, M. W., Crone, E. A., … Grafton, S. (2012). Social exclusion and punishment of excluders: Neural correlates and developmental trajectories. *NeuroImage, 59*(1), 708–717. Retrieved from https://www.sciencedirect.com/science/article/pii/S1053811911007890

Kross, E., Berman, M. G., Mischel, W., Smith, E. E., & Wager, T. D. (2011). Social rejection shares somatosensory representations with physical pain. *Proceedings of the National Academy of Sciences, 108*(15), 6270–6275. https://doi.org/10.1073/pnas.1102693108

Lee, H., Ku, J., Kim, J., Jang, D.-P., Yoon, K. J., Kim, S. I., & Kim, J.-J. (2014). Aberrant neural responses to social rejection in patients with schizophrenia. *Social Neuroscience, 9*(4), 412–423. https://doi.org/10.1080/17470919.2014.907202

Lieberman, M. D. (2010). Social cognitive neuroscience. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 143–193). New York, NY: McGraw-Hill.

Lieberman, M. D., Straccia, M. A., Meyer, M. L., Du, M., & Tan, K. M. (2019). Social, self, (situational), and affective processes in medial prefrontal cortex (MPFC): Causal, multivariate, and reverse inference evidence. *Neuroscience & Biobehavioral Reviews, 99*, 311. https://doi.org/10.1016/J.NEUBIOREV.2018.12.021

Lin, N. (2002). *Social capital: A theory of social structure and action*. Retrieved from https://books.google.com/books?hl=en&lr=&id=fvBzIu5-yuMC&oi=fnd&pg=PR11&dq=Lin+N+(2002)+Social+Capital:+A+Theory+of+Social+Structure+and+Action+(Cambridge+Univ+Press,&ots=UW-Di_wHAX&sig=6AkGq8fNfELQ6d_I3h-laoCH30Q

Liu, D., Sabbagh, M. A., Gehring, W. J., & Wellman, H. M. (2009). Neural correlates of children's theory of mind development. *Child Development, 80*(2), 318–326. https://doi.org/10.1111/j.1467-8624.2009.01262.x

MacDonald, G., & Leary, M. R. (2005). Why does social exclusion hurt? The relationship between social and physical pain. *Psychological Bulletin, 131*, 202. https://doi.org/10.1037/0033-2909.131.2.202

Masten, C. L., Colich, N. L., Rudie, J. D., Bookheimer, S. Y., Eisenberger, N. I., & Dapretto, M. (2011). An fMRI investigation of responses to peer rejection in adolescents with autism spectrum disorders. *Developmental Cognitive Neuroscience, 1*(3), 260–270. https://doi.org/10.1016/J.DCN.2011.01.004

McPartland, J. C., Crowley, M. J., Perszyk, D. R., Naples, A. J., Mukerji, C. E., Wu, J., … Mayes, L. C. (2011). Temporal dynamics reveal atypical brain response to social exclusion in autism. *Developmental Cognitive Neuroscience, 1*(3), 271–279. https://doi.org/10.1016/j.dcn.2011.02.003

Mier, D., Sauer, C., Lis, S., Esslinger, C., Wilhelm, J., Gallhofer, B., & Kirsch, P. (2010). Neuronal correlates of affective theory of mind in schizophrenia out-patients: Evidence for a baseline deficit. *Psychological Medicine, 40*(10), 1607–1617. https://doi.org/10.1017/S0033291709992133

Moor, B. G., van Leijenhorst, L., Rombouts, S. A. R. B., Crone, E. A., & van der Molen, M. W. (2010). Do you like me? Neural correlates of social evaluation and developmental trajectories. *Social Neuroscience, 5*(5), 461–482. https://doi.org/10.1080/17470910903526155

Muscatell, K. A., Morelli, S. A., Falk, E. B., Way, B. M., Pfeifer, J. H., Galinsky, A. D., … Eisenberger, N. I. (2012). Social status modulates neural activity in the mentalizing network. *NeuroImage, 60*(3), 1771–1777. https://doi.org/10.1016/j.neuroimage.2012.01.080

Onoda, K., Okamoto, Y., Nakashima, K., Nittono, H., Ura, M., & Yamawaki, S. (2009). Decreased ventral anterior cingulate cortex activity is associated with reduced social pain during emotional support. *Social Neuroscience, 4*(5), 443–454. https://doi.org/10.1080/17470910902955884

Panksepp, J. (2004). *Affective neuroscience: The foundations of human and animal emotions*. Oxford, England: Oxford University Press.

Park, I. H., Ku, J., Lee, H., Kim, S. Y., Kim, S. I., Yoon, K. J., & Kim, J. J. (2011). Disrupted theory of mind network processing in response to idea of reference evocation in schizophrenia. *Acta Psychiatrica Scandinavica, 123*(1), 43–54. https://doi.org/10.1111/j.1600-0447.2010.01597.x

Rochat, P. (2003). Five levels of self-awareness as they unfold early in life. *Consciousness and Cognition, 12*, 717–731. https://doi.org/10.1016/S1053-8100(03)00081-3

Rochat, P. (2009). Others in mind: Social origins of self-consciousness. Cambridge University Press.

Sabbagh, M. A., Bowman, L. C., Evraire, L. E., & Ito, J. M. B. (2009). Neurodevelopmental correlates of theory of mind in preschool children. *Child Development, 80*(4), 1147–1162. https://doi.org/10.1111/j.1467-8624.2009.01322.x

Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology, 16*(2), 235–239. https://doi.org/10.1016/j.conb.2006.03.001

Schmälzle, R., Brook O'Donnell, M., Garcia, J. O., Cascio, C. N., Bayer, J., Bassett, D. S., … Falk, E. B. (2017). Brain connectivity dynamics during social interaction reflect social network structure. *Proceedings of the National Academy of Sciences, 114*(20), 5153–5158. https://doi.org/10.1073/pnas.1616130114

Sebastian, C., Blakemore, S. J., & Charman, T. (2009). Reactions to ostracism in adolescents with autism spectrum conditions. *Journal of Autism and Developmental Disorders, 39*(8), 1122–1130. https://doi.org/10.1007/s10803-009-0725-4

Sebastian, C., Viding, E., Williams, K. D., & Blakemore, S. J. (2010). Social brain development and the affective consequences of ostracism in adolescence. *Brain and Cognition, 72*(1), 134–145. Retrieved from https://www.sciencedirect.com/science/article/pii/S0278262609001055

Sebastian, C. L., Tan, G. C. Y., Roiser, J. P., Viding, E., Dumontheil, I., & Blakemore, S.-J. (2011). Developmental influences on the neural bases of responses to social rejection: Implications of social neuroscience for education. *NeuroImage, 57*(3), 686–694. Retrieved from https://www.sciencedirect.com/science/article/pii/S1053811910012656

Silk, J. S., Stroud, L. R., Siegle, G. J., Dahl, R. E., Lee, K. H., & Nelson, E. E. (2012). Peer acceptance and rejection through the eyes of youth: Pupillary, eyetracking and ecological data from the chatroom interact task. *Social Cognitive and Affective Neuroscience, 7*, 93. Retrieved from https://academic.oup.com/scan/article-abstract/7/1/93/1638951

Somerville, L. H. (2013). The teenage brain: Sensitivity to social evaluation. *Current Directions in Psychological Science, 22*, 121. https://doi.org/10.1177/0963721413476512

Somerville, L. H., Heatherton, T. F., & Kelley, W. M. (2006). Anterior cingulate cortex responds differentially to expectancy violation and social rejection. *Nature Neuroscience, 9*, 1007–1008. Retrieved from https://www.nature.com/articles/nn1728

Spunt, R. P., Falk, E. B., & Lieberman, M. D. (2010). Dissociable neural systems support retrieval of how and why action knowledge. *Psychological Science, 21*(11), 1593–1598. https://doi.org/10.1177/0956797610386618

Vijayakumar, N., Cheng, T. W., & Pfeifer, J. H. (2017). Neural correlates of social exclusion across ages: A coordinate-based meta-analysis of functional MRI studies. *NeuroImage, 153*, 359–368. https://doi.org/10.1016/j.neuroimage.2017.02.050

Wang, J., Iannotti, R. J., & Nansel, T. R. (2009). School bullying among adolescents in the United States: Physical, verbal, relational, and cyber. *Journal of Adolescent Health, 45*(4), 368–375. https://doi.org/10.1016/j.jadohealth.2009.03.021

Wegner, D. M., & Gray, K. (2008). The sting of intentional pain. *Psychological Science, 19*, 1260–1262. https://doi.org/10.1088/0960-1317/20/10/104002

Williams, K. D., Cheung, C. K., & Choi, W. (2000). Cyberostracism: Effects of being ignored over the internet. *Journal of Personality and Social Psychology, 79*(5), 748–762. Retrieved from http://psycnet.apa.org/journals/psp/79/5/748.html?uid=2000-00920-006

Zadro, L., Williams, K. D., & Richardson, R. (2004). How low can you go? Ostracism by a computer is sufficient to lower self-reported levels of belonging, control, self-esteem, and meaningful existence. *Journal of Experimental Social Psychology, 40*, 560–567. https://doi.org/10.1016/j.jesp.2003.11.006

Putting the "Me" in "Mentalizing": Multiple Constructs Describing *Self* Versus *Other* During Mentalizing and Implications for Social Anxiety Disorder

Erin L. Maresh and Jessica R. Andrews-Hanna

Introduction

In daily life, the experience of reflecting on our own thoughts and feelings may subjectively feel quite distinct from the experience of inferring the thoughts and feelings of other people. Yet, it is becoming increasingly appreciated that the processes underlying how we understand the mental states of both ourselves and others—processes collectively called "mentalizing"—show considerable overlap and interconnectedness (Gerace, Day, Casey, & Mohr, 2017; Oosterwijk, Snoek, Rotteveel, Barrett, & Steven Scholte, 2017; Saxe, 2015). For example, reflecting on our own thoughts, feelings, and memories may provide a template for understanding the mental states of others (Bradford, Jentzsch, & Gomez, 2015; Dimaggio, Lysaker, Carcione, Nicolò, & Semerari, 2008; Gordon, 1986; van der Meer, Costafreda, Aleman, & David, 2010). Inversely, attempting to understand others' mental states can clarify our own inner experience and self-concept (Cooley, 1909; Fonagy, Gergely, Jurist, & Target, 2002; Mead, 1934). Consequently, far from distinct constructs, self- and other-mentalizing are interdependent processes with broad implications for psychopathology, where both excessive and limited self-focus can be associated with impairments in understanding others (Cotter et al., 2018; Dimaggio et al., 2008; Kaplan et al., 2018; Plana, Lavoie, Battaglia, & Achim, 2014). To date, however, self-focused thought has been explored largely independently from mentalizing about others, and hence, little is known about how self-focus benefits or impairs mentalizing.

E. L. Maresh (✉)
Department of Psychology, University of Arizona, Tucson, AZ, USA
e-mail: erinmaresh@arizona.edu

J. R. Andrews-Hanna (✉)
Department of Psychology, Cognitive Science Program, University of Arizona,
Tucson, AZ, USA
e-mail: jandrewshanna@arizona.edu

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_33

The aim of this chapter is to begin refining our understanding of the relationship between the *self* and *other*. Specifically, we will examine different ways of understanding the role of the self in mentalizing and consider its relevance to social anxiety disorder (SAD). To this end, we will establish three distinct but overlapping constructs describing different ways of construing *self* versus *other* in mentalizing. For each construct, we will integrate behavioral and neural evidence from diverse fields, highlighting a critical role for the brain's default mode network (DMN) in supporting these constructs (Andrews-Hanna, Smallwood, & Spreng, 2014; Mars et al., 2012; Northoff et al., 2006; Spreng & Andrews-Hanna, 2015), and will discuss how heightened focus on the *self* within each construct contributes to SAD. SAD, a disorder characterized by excessive fear of being evaluated by others, is hypothesized to be maintained by negative self-focused thought related to social situations (Alden, Auyeung, & Plasencia, 2014; Heimberg, Brozovich, & Rapee, 2010), making it especially suited to examining how self-focus interferes with mentalizing about others. Finally, we will consider real-world examples of these constructs and broader clinical implications. Our hope is that by shedding light on the interdependence of self- and other-processing in mentalizing, we will inform our understanding of both functional and dysfunctional mentalizing, uncover potential transdiagnostic targets for therapeutic intervention, and highlight exciting areas for future research.

Constructs to Distinguish the Self and Other

Even the simplest social exchange engages a complex interplay between processing the self and processing others. We can flexibly switch between considering our own mental states and those of our interaction partners; we can infer the emotions and perspectives of others without confusing them with our own; and we can dynamically evoke mental images of ourselves and of others in past and potential future scenarios to inform our social behavior. Thus, far from a singular construct, distinguishing between the self and other during mentalizing likely involves multiple underlying constructs. In the sections that follow, we describe three of these constructs, with an emphasis on the role of the *self* in each: (1) when the self is the *target* of mental state inferences, (2) when the self is the *source* of knowledge used to make mental state inferences, and (3) when an image of the self is mentally constructed due to the *visual perspective* adopted during mental imagery. Of note, throughout this chapter, we use the term "mentalizing" to indicate making mental state inferences not only about *cognitive* states, such as thoughts, beliefs, and intentions, but also about *affective* states, given the interdependence of neural processes underlying cognitive and affective mentalizing (Lamm & Majdandžić, 2015; Sebastian et al., 2012).

Self as Target: Understanding One's Own Mental State

Mentalizing is often construed as the ability to infer the mental states of other people (Frith & Frith, 2006; Mitchell, 2006). Yet, equally important to its definition is the ability to infer one's *own* mental states, a process that has been referred to by many names, including "self-referential thought," "self-reflection," "private self-consciousness," and—harkening back to William James—"introspection" (Fenigstein, Scheier, & Buss, 1975; James, 1890; van der Meer et al., 2010). Here, we will call this process "self-focused mentalizing" to underscore its role in mentalizing while differentiating it from *other*-focused mentalizing. Thus, perhaps the most overt construct for distinguishing "self" and "other" in mentalizing is simply identifying the *target* of mental state inference—whether the *perceiver* (the individual making a mental state attribution) is trying to understand their own mental state (self-focused mentalizing) or that of another person (other-focused mentalizing).

Although identifying *self* or *other* as the target of mentalizing appears straightforward at first glance, several methodological issues hamper its precise determination. Various methods for constraining the target of mentalizing include varying task *content* (e.g., whether the task contains stimuli relevant to the self or to another), task *context* (e.g., whether the task is performed alone or with others), or task *instructions* (e.g., whether the perceiver is told to think about their own thoughts and feelings or those of another). However, these techniques rely on assumptions that are difficult to establish, including (1) that self-relevant stimuli and solitary tasks produce only self-focused mentalizing, and other-relevant stimuli and interactive tasks (e.g., trust games) produce only other-focused mentalizing, and (2) that the target of mentalizing remains static in a situation rather than, for example, dynamically shifting between the self and other(s). While these assumptions may hold true in simplified and contrived task designs, they are unlikely to maintain during complex, *naturalistic* instances of social cognition (Zaki & Ochsner, 2009). As such, little is known about natural variation in the degree to which individuals actually mentalize about themselves or others (but see Bryant, Coffey, Povinelli, & Pruett, 2013) or how "target-switching" might dynamically unfold during a social interaction.

Adding to the difficulty in determining the target of mentalizing is the question of whether and how the processes underlying mentalizing about the *self* differ from the processes underlying mentalizing about *others* (e.g., Legrand & Ruby, 2009). It has been suggested that, during self-focused mentalizing, we have access to multiple facets of our inner experience, such as physiological states, affective reactions, and memories (Damasio, 2010; Varela, Thompson, & Rosch, 2017), that provide privileged information about ourselves not available when mentalizing about others. Further, it intuitively *feels* like we know ourselves better than anyone else. Despite these intuitions, we are prone to significant self-perception biases that limit our self-knowledge, including the suggestion that many facets of personal experience occur largely outside of conscious awareness and thus cannot be readily accessed for the purposes of mentalizing (Vazire & Carlson, 2010; Wilson & Dunn, 2004).

Self-focused mentalizing may, instead, occur primarily through a constructive process, operating similarly to how we are believed to understand others. That is, we may use observations of our behaviors and reactions (rather than introspective processes) to make inferences about our mental states and then construct a personal narrative from these inferences (Bem, 1972; Bollich, Johannet, & Vazire, 2011; Wilson & Dunn, 2004). Supporting this idea, it has been proposed that self-focused mentalizing can be improved by seeking out information from others, both through observing other people's reactions to one's own actions and through exploring other people's differing views (Bollich et al., 2011; Wilson & Dunn, 2004). In other words, seeking out *other*-focused mentalizing may be critical in improving *self*-focused mentalizing, highlighting the interdependence of self and other processes in mental state inference.

Neural correlates. Numerous studies have sought to identify the neural correlates of self-focused mentalizing and other self-related processes, reliably identifying activity within the core structures of the DMN, including the medial prefrontal cortex (mPFC), posterior cingulate cortex (PCC), and anterior cingulate cortex (ACC) (Andrews-Hanna et al., 2014; Northoff et al., 2006; Qin & Northoff, 2011; van der Meer et al., 2010). Yet, recent evidence suggests that regions involved in mentalizing about the self overlap with many regions involved in mentalizing about others, raising the question of what, if any, activation in the brain is self-specific (Legrand & Ruby, 2009; Qin & Northoff, 2011; van der Meer et al., 2010).

Within mentalizing research, particular attention has been given to the mPFC (Denny, Kober, Wager, & Ochsner, 2012; Schilbach, Eickhoff, Rotarska-Jagiela, Fink, & Vogeley, 2008; Spreng & Andrews-Hanna, 2015), especially for its hypothesized role in distinguishing between self and other. Specifically, the mPFC has been theorized to map representations of *self* and *other* along a spatial gradient, with more ventral mPFC portions proposed to predominately represent the *self*, and more dorsal mPFC proposed to predominately represent *others* (Denny et al., 2012; Lieberman, Straccia, Meyer, Du, & Tan, 2019; van der Meer et al., 2010).

Supporting this distinction, more ventral portions of the mPFC are involved in a range of processes related to the self, including encoding and prioritizing self-relevant information in memory (Kumaran, Banino, Blundell, Hassabis, & Dayan, 2016; Macrae, Moran, Heatherton, Banfield, & Kelley, 2004), retrieving autobiographical facts and episodes (Svoboda, McKinnon, & Levine, 2006), referencing information to one's self (Northoff et al., 2006), and constructing personal meaning from stimuli (Roy, Shohamy, & Wager, 2012). However, brain activity in ventral mPFC regions has been found to track not only the degree of self-relatedness of a stimulus but also its perceived value (Andrews-Hanna, Reidler, Sepulcre, Poulin, & Buckner, 2010; Bartra, McGuire, & Kable, 2013; Moran, Heatherton, & Kelley, 2009), with recent pattern-based neuroimaging studies suggesting at least partial overlap of these two processes at the representational level (Chavez, Heatherton, & Wagner, 2017; Yankouskaya et al., 2017). Ventral mPFC regions may therefore play a broader role in computing the personal significance or motivational salience of external or internal information, rather than processing self-relatedness per se (Andrews-Hanna et al., 2014; D’Argembeau, 2013; Moran et al., 2009). In line with this notion, ventral portions of the mPFC become engaged to a greater degree when mentalizing about psychologically close or similar others, as compared to strangers or dissimilar others (Krienen, Tu, & Buckner, 2010; Mitchell, Macrae, & Banaji, 2006; Murray, Schaer, & Debbané, 2012; Tamir & Mitchell, 2010).

In contrast to its ventral portions, activation in dorsal mPFC (dmPFC) is often observed during tasks that involve *other*-focused mentalizing, including theory of mind paradigms and other controlled or *reflective* (as opposed to automatic or *reflexive*) social cognitive tasks (Lieberman, 2007; Saxe, 2015). Within the DMN, the dmPFC, along with the inferior frontal gyrus (IFG), temporoparietal junction (TPJ), superior temporal sulcus, and temporal poles, is thought to form a functionally coherent “dmPFC subsystem” (Andrews-Hanna et al., 2014; Yeo et al., 2011; but see Braga & Buckner, 2017) that strongly overlaps with several regions of the “mentalizing network” (Spreng & Andrews-Hanna, 2015). Despite evidence for preferential activity within the dmPFC subsystem for other-focused mentalizing, many of these regions are also recruited when mentalizing about the self, particularly when making reflective self-focused inferences (Denny et al., 2012). Further, a growing body of research has begun to highlight the role of the dmPFC and other regions in the subsystem in high-level non-social processes involving abstract construals (Baetens, Ma, Steen, & Van Overwalle, 2013; Baetens, Ma, & Van Overwalle, 2017) and narrative comprehension (Mar, 2010; Tamir, Bricker, Dodell-Feder, & Mitchell, 2015). This suggests that activity in the dmPFC is not specific to other-focused mentalizing, paralleling findings regarding ventral mPFC activity and self-focused mentalizing.

Given these alternative accounts of their function, ventral and dorsal subregions of the mPFC have been proposed to be “agent-independent”—that is, they do not inherently distinguish between representations of self and other but rather encode qualities that often *correspond* with differences between self and other, such as abstraction, subjective value, relevance (e.g., information related to the self is more likely to be experienced as concrete, valuable, and relevant) (Garvert, Moutoussis, Kurth-Nelson, Behrens, & Dolan, 2015; Nicolle et al., 2012). We suggest that the dmPFC subsystem plays an important role in both other-focused *and* self-focused mentalizing, particularly when processes involve conceptually abstract, reflective mental processes represented verbally or symbolically (Gilead, Trope, & Liberman, 2019; Raffaelli, Wilcox, & Andrews-Hanna, 2020).

Relevance to social anxiety. Excessive and maladaptive self-focused mentalizing is thought to be critical to the generation and maintenance of SAD (Alden, Auyeung, & Plasencia, 2014; Heimberg et al., 2010). During social situations, individuals with SAD are hypothesized to focus their attention on themselves, monitoring their own thoughts, feelings, and internal sensations to form an image of how others might be seeing them, rather than on social or environmental cues (Heimberg et al., 2010; Maresh, Allen, & Coan, 2014; Maresh, Teachman, & Coan, 2017).

We hypothesize that, in addition to exacerbating social anxiety and other negative outcomes, excessive self-focus in SAD likely interferes with mentalizing about others. Surprisingly little work has examined other-focused mentalizing in SAD, despite ample research linking SAD with interpersonal difficulties (reviewed in Alden, Regambal, & Plasencia, 2014). We posit at least three ways that self-focused mentalizing in SAD might interfere with other-focused mentalizing: (1) by diverting limited attentional resources away from understanding the other and toward monitoring the self (Eysenck & Derakshan, 2011), (2) by shifting other-focused mentalizing to be about self-relevant information (i.e., *reflected self-appraisals*; Wallace & Tice, 2012), and (3) by facilitating avoidance behaviors, such as eye gaze avoidance or restricted speech, that are intended to reduce anxiety but also prevent attending to information about one's conversation partner (Plasencia, Alden, & Taylor, 2011). Thus, during social situations, in which a focus on understanding the mental states of others is critical, individuals with SAD may be focusing on "self-monitoring," spending considerable effort analyzing social interactions for self-referential cues, and restricting social behaviors at the expense of gathering accurate other-focused information.

Individuals with SAD may spend less time engaging in other-focused mentalizing due to heightened attention to the self, but how do they perform when they *are* mentalizing about others? While many studies suggest that social anxiety confers impairments in other-focused mentalizing, mixed results indicate a more complicated picture. Individuals higher in social anxiety report lower tendencies toward taking others' perspectives (Beitel, Ferrer, & Cecero, 2005; Davis, 1983; Davis & Franzoi, 1991), paralleled by poorer performance on perspective-taking tasks compared to their non-anxious counterparts (Buhlmann, Wacker, & Dziobek, 2015; Hezel & McNally, 2014; Lenton-Brym, Moscovitch, Vidovic, Nilsen, & Friedman, 2018; Washburn, Wilson, Roes, Rnic, & Harkness, 2016). When specific *types* of mentalizing errors are assessed, individuals with SAD make errors indicative of *over*-mentalizing (Hezel & McNally, 2014; Washburn et al., 2016)—that is, reading too much into what others are thinking and feeling. In addition to over-mentalizing, socially anxious individuals may be more likely to inaccurately infer that others' thoughts are focused on them, evaluating their appearance and/or performance (Hope, Burns, Hayes, Herbert, & Warner, 2010; Stopa & Clark, 1993).

Although the majority of studies find that social anxiety corresponds with impairments in other-focused mentalizing, some studies have found that individuals high in social anxiety exhibit *superior* other-focused mentalizing—at least during certain tasks and under certain circumstances. For example, socially anxious individuals under social-evaluative threat were more accurate at evaluating others' negative emotions (Auyeung & Alden, 2016), and socially anxious women (but not men) were more accurate at gauging whether another person was lying (Sutterby, Bedwell, Passler, Deptula, & Mesa, 2012). Other studies find no differences in other-focused mentalizing abilities related to SAD (Morrison et al., 2016). Due to the wide variety of methods, sample characteristics, and analytic approaches used in these studies, whether SAD interferes with other-focused mentalizing, and whether excessive self-focus is a key mechanism in this interference, remain important avenues of future research.

Neurally, individuals with SAD, relative to healthy controls, show aberrant processing of self-referential stimuli across regions of the DMN—including the mPFC. SAD corresponds with heightened activity in ventral mPFC during a broad array of self-referential tasks regardless of stimulus valence, including viewing self-referential words (Blair et al., 2008), receiving social feedback (Peterburs, Sandrock, Miltner, & Straube, 2016), and viewing second-person compared to first-person self-referential statements (Blair et al., 2011). Interestingly, SAD also corresponds with heightened activity in the dmPFC during self-referential tasks—but predominately in response to *negative* stimuli, such as receiving negative criticism (Blair et al., 2008), viewing social anxiety-related scenes/words (Heitmann et al. 2016, 2017), anticipating unpleasant stimuli (Brühl et al., 2011), viewing distracting emotional faces (Boehme et al., 2015), and experiencing public embarrassment (Müller-Pinzler et al., 2015). Rather than encoding differences in self and other, increased ventral mPFC activation may indicate greater salience of general self-related stimuli in SAD relative to non-socially anxious individuals, whereas increased dmPFC activation—seen specifically during processing of negative self-referential stimuli—may support heightened abstract, narrative, and ruminative processes adopted during negative self-focused thought in SAD (Andrews-Hanna, Christoff, & O'Connor, 2020).

Self as Source: Egocentricity in Mentalizing Representations

Regardless of whether the *target* of mentalizing is the self or another, the individual serving as the primary *source* from which mental state inferences are drawn can also be either the self or another. It has been proposed that successful mentalizing involves not only accurately inferring the target's mental state, but also inhibiting irrelevant perspectives—that is, one's own perspective if mentalizing about others, and others' perspectives if mentalizing about oneself (Leslie, Friedman, & German, 2004; Samson, Apperly, Kathirgamanathan, & Humphreys, 2005). For various reasons, however, we sometimes fail to inhibit irrelevant perspectives when mentalizing, leading to either *egocentric biases,* in which our own beliefs become the source of our inferences about others (Peters, 2016), or *altercentric biases,* in which another person's beliefs become the source of our self-inferences (De Vignemont & Mercier, 2016). Here, we will focus on egocentric biases, in which the *target* of mentalizing is another person, but the *source* of mentalizing is the self.

Relying on one's own mental states to understand another can be problematic across multiple circumstances, resulting in erroneous self-attributions onto the target (Keysar, Barr, Balin, & Brauner, 2000; Steinbeis & Singer, 2014). For example, inaccurate egocentric biases are more likely to occur when the perceiver has different traits than the target (Krueger & Clement, 1994), experiences a different affective response than the target (Steinbeis & Singer, 2014), or possesses privileged information that is unknown to the target (Apperly, Back, Samson, & France, 2008). To overcome egocentricity biases, an *anchor-and-adjust* approach may be deployed in which inferences about another person's perspective are first egocentrically "anchored" in self-knowledge and are then "adjusted" according to known and estimated discrepancies between the self and other (Epley, Keysar, Van Boven, & Gilovich, 2004; Tamir & Mitchell, 2010, 2013). Although largely beneficial, the anchor-and-adjust approach has limitations. Chiefly, it is constrained by an individual's preexisting knowledge about the other person—if you know little relevant information about someone, there will be little adjusting you can do. Further, adjusting from egocentric self-knowledge is cognitively demanding, with greater perceived differences between *self* and *other* requiring more effortful, stepwise cognitive processing. To reduce effort when mentalizing about a dissimilar other, perceivers may anchor their mental state inferences in a familiar, well-known person (e.g., a significant other) instead of their own self-knowledge. This anchoring in another appears to occur primarily when the familiar other is a better exemplar than the self of the aspect being assessed in the target person (Willard & Markman, 2017).

Given that overcoming egocentricity is an effortful process, what determines whether we remain with our default egocentric biases or put forth effort to adjust our attributions? Sufficient time and motivation can increase the likelihood of anchoring-and-adjustment (Epley et al., 2004; Stern & West, 2016). However, even when engaged in anchoring-and-adjustment, a perceiver may cease making effortful adjustments prematurely, settling on a "satisfactory" estimate of the other person's mental state based on the amount of effort the perceiver is willing to expend (Epley & Gilovich, 2006). We are more likely to effortfully anchor-and-adjust with those who are similar to us than with those who are dissimilar to us, at least with unfamiliar others (Tamir & Mitchell, 2013). Rather than using egocentric biases to understand dissimilar others, however, we may instead rely on stereotypes (Ames, 2004), possibly because adjusting to the perspective of someone unlike us is deemed too effortful. Surprisingly, with familiar others, such as friends or spouses, we are more likely to rely on egocentric biases than to adjust our inferences (Savitsky, Keysar, Epley, Carter, & Swanson, 2011), suggesting that we overestimate the degree to which close others share our perspectives. Future research may wish to examine how familiarity and similarity interact to influence egocentricity, given their seemingly divergent effects on whether self or other is the source of mental state inference.

Although much research casts egocentricity in a negative light, egocentric inference can provide a useful heuristic in which readily accessible self-knowledge is used to gain insight into others' experiences (Hoch, 1987; Keysar et al., 2000). Further, making egocentric inferences is related to reduced stereotyping, increased prosocial behavior, and greater intimacy, suggesting that egocentric inferences may, in some cases, be tied to viewing others more like the self (Coan & Maresh, 2014; Galinsky, Ku, & Wang, 2005)—a process conceptually similar to "self-other overlap" (Aron, Lewandowski, Mashek, & Aron, 2013). Perhaps most importantly, egocentric inferences can be accurate when applied to people who are similar to us, allowing us to save resources when applied judiciously (Ames, 2004; Ames, Weber, & Zou, 2012; Hoch, 1987; Keysar et al., 2000). It is important to note that egocentric inferences are usually only identified as egocentric *because* they are inaccurate; accurate egocentric inferences are more difficult to identify (Wallin, 2011). As such, although literature has emphasized the role of egocentricity in mental state attribution *errors*, it is possible that egocentric attributions are, in general, more accurate than the literature would suggest (Eyal, Steffel, & Epley, 2018; Keysar et al., 2000).

Neural correlates. Inhibiting egocentric mental states when mentalizing about another is thought to be related to a broader ability to flexibly switch between representations of self and other (de Guzman, Bird, Banissy, & Catmur, 2016). Control of self-other representations is supported by regions implicated in general inhibitory control, such as the IFG and dorsolateral PFC (Hartwright, Apperly, & Hansen, 2012; Rothmayr et al., 2011; Van der Meer, Groenewold, Nolen, Pijnenborg, & Aleman, 2011), as well as two regions of the DMN found in right temporoparietal cortex—the TPJ and supramarginal gyrus (SMG)—that show differential control over cognitive and affective states (de Guzman et al., 2016; Silani, Lamm, Ruff, & Singer, 2013; Steinbeis, 2016). The right TPJ (rTPJ), particularly the posterior subregion (Igelstrom, Webb, & Graziano, 2015), contributes to inhibiting cognitive mental states, including beliefs (Hartwright et al., 2012; Rothmayr et al., 2011; Van der Meer et al., 2011) and visual perspectives (Santiesteban, Banissy, Catmur, & Bird, 2012). Inhibiting affective mental states, on the other hand, appears related to activation of the right SMG (rSMG), which lies anterior to the rTPJ (Silani et al., 2013; Steinbeis, Bernhardt, & Singer, 2015). Supporting their roles in different aspects of perspective inhibition, rTPJ and rSMG show distinct connectivity profiles, with posterior rTPJ coupling with other regions implicated in cognitive mental state attribution, such as the PCC, precuneus, and mPFC, and rSMG coupling with regions related to affective empathy, including the midcingulate cortex and anterior insula (Hoffmann, Koehne, Steinbeis, Dziobek, & Singer, 2016; Steinbeis et al., 2015).

While a number of studies have examined the neural correlates of *inhibiting* egocentric inferences, relatively few have examined neural correlates of what underlies egocentric inferences in the first place. Egocentricity biases may be partially rooted in, and/or influenced by, *shared representations* of mental states—overlapping neural activity seen both when experiencing (or imagining) a mental state and when interpreting another as experiencing the same mental state (Ochsner et al., 2008; Oosterwijk et al., 2017; Steinbeis & Singer, 2014). Shared representations may be inherently egocentric because they ultimately rely on not only our idiosyncratic patterns of neural activity when representing a given mental state but also our personal interpretation of what mental state an individual is likely to be experiencing (Lamm, Bukowski, & Silani, 2016); as such, shared representations are thought to include activity in self-related core DMN regions, such as the mPFC, precuneus/PCC, and ACC (Lombardo et al., 2010; Northoff et al., 2006; van der Meer et al., 2010). Greater egocentricity when judging others' emotions has also been linked with reduced recruitment of the rSMG and reduced coupling between the rSMG and dlPFC (Steinbeis et al., 2015).

The process of adjusting from egocentric inferences to adopt another person's perspective is linked to activity in the dmPFC, which shows a linear increase in activation with greater perceived discrepancy between the self and other (Tamir & Mitchell, 2010). While this may support the hypothesis that the dmPFC is specific to other-focused mentalizing, we believe it provides more compelling evidence for the role of the dmPFC in reflective, abstract thought more broadly, as attempting to understand someone—especially someone very different from oneself—likely involves high-level construal (Koster-Hale et al., 2017). More ventral regions of the mPFC also show increased activity in response to perceived discrepancies between self and other, but, unlike the dmPFC, this does not vary based on the extent of the discrepancy. Following an anchor-and-adjust model, activity in more ventral mPFC may represent initial anchoring in self-knowledge, whereas the dmPFC underlies the subsequent adjustment process (Tamir & Mitchell, 2010).

Relevance to social anxiety. Although the role of the self as a *target* of mentalizing in SAD is well established, little research has explicitly examined the role of the self as a *source* of mentalizing in SAD. However, cognitive models suggest that socially anxious people rely on egocentric inferences to understand others' mental states, particularly in relation to reflected self-appraisals (Clark & Wells, 1995; Heimberg et al., 2010). Individuals with SAD create mental images of themselves during social situations that are purported to model the perceptions of *others;* yet, these images appear to be composed predominately of negative internal self-perspectives, including distorted self-schemas, images of past distressing social situations, and current physiological manifestations of anxiety (Hope et al., 2010; Stopa & Clark, 1993). In other words, the experience of social anxiety may evoke egocentricity biases in which the socially anxious individual's negative self-appraisal is used as a template for others' perceptions of the self.

Although speculative, some evidence suggests that individuals with SAD also show egocentricity biases more generally, in the absence of self-referential stimuli. For example, individuals with SAD display over-mentalizing errors when making mental state inferences about characters acting in a film (Hezel & McNally, 2014; Washburn et al., 2016), indicating that over-mentalizing in SAD is not necessarily tied to self-focused fears, such as searching others for signs of negative evaluation. Rather, over-mentalizing may be evidence of a general egocentricity bias, such that those with SAD, who experience more negative emotions and beliefs than healthy individuals (Gros & Sarver, 2014), project these emotions and beliefs onto others. However, whether SAD corresponds with difficulty inhibiting one's self-perspective during mentalizing is largely untested.

Some evidence suggests that egocentricity bias in social anxiety offers some benefits. In general, people tend to underestimate others' pain, both physical and social (Kappesser, Williams, & Prkachin, 2006; Nordgren, Banas, & MacDonald, 2011). Interestingly, individuals higher in social anxiety showed less underestimation—indicating better accuracy—when mentalizing about others' negative social emotions, but only while they were under social-evaluative threat (Auyeung & Alden, 2016). It is possible that socially anxious participants experienced heightened negative affect while being evaluated, which they projected onto others (Todd, Forstmann, Burgmer, Brooks, & Galinsky, 2015)—ultimately resulting in less underestimation of others' negative affect. Although speculative, this could be further evidence that, under certain circumstances, egocentricity biases can be beneficial. It should be noted that in this study, the link between social anxiety and improved accuracy for negative emotions was identified in a *non-selected* sample—that is, participants were not recruited based on levels of social anxiety, and no diagnostic information was collected (Auyeung & Alden, 2016). Thus, it is unknown how clinical levels of social anxiety might interact with social-evaluative threat to impact mentalizing.

Self as Object: Visual Perspective During Mentalizing-Related Imagery

Through mental imagery, we are able to engage in "mental time travel" in which we construct internal representations of the past, present, or future that can be derived from actual events (e.g., autobiographical memory) or imagined ones (e.g., future episodic thinking, counterfactual thinking) (Moulton & Kosslyn, 2009; Schacter, Benoit, De Brigard, & Szpunar, 2015; Suddendorf, Addis, & Corballis, 2009). Simulating events via mental imagery is hypothesized to serve many social cognitive functions, allowing us to generate predictions to guide future social behavior, rehearse responses to upcoming interactions, review an interaction after it has taken place, and reflect on our own thoughts and feelings following an interaction (Bar, 2009; Honeycutt & Ford, 2006; Libby & Eibach, 2013; Moulton & Kosslyn, 2009). One important aspect of mental imagery likely to influence mentalizing is the *visual perspective* used to picture it. Individuals can perceive mental imagery as if through their own eyes, called the *field* or *first-person perspective*, or as if through the eyes of a spectator observing the scene, called the *observer* or *third-person perspective* (Libby & Eibach, 2011; Nigro & Neisser, 1983; Sutin & Robins, 2008).[1]

The visual perspective adopted during mental imagery impacts several phenomenological features of the imagined event. When using a first-person perspective, our field of view more closely resembles how we visually perceive the world in "real life"—embodied within our imagined self, acting as the *subject* of the scene. Accordingly, compared to third-person imagery, first-person imagery tends to be more visually vivid (Butler, Rice, Wooldridge, & Rubin, 2016) and more physically and emotionally intense (Holmes, Coughtrey, & Connor, 2008; McIsaac & Eich, 2002; Pronin & Ross, 2006). In contrast, adopting a third-person perspective

[1] We acknowledge that the term "perspective" has many meanings, particularly in mentalizing research. In this section, we will use the term "perspective" solely to indicate the *visual viewpoint* adopted during mental imagery and not to indicate the concept of adopting another person's mental state in the here and now as it is used in psychological perspective-taking literature (e.g., Galinsky et al., 2005).

requires constructing a visual image of the self, such that the imagined self is perceived as an *object* of mental imagery, rather than the embodied, agentic *subject* (Libby & Eibach, 2011; Sutin & Robins, 2008). Memories recalled from a third-person perspective de-emphasize visual detail and affective salience, instead tending to focus on the "actors" in the scene and evaluating their traits, behaviors, and appearance (Libby, Valenti, Pfent, & Eibach, 2011; McIsaac & Eich, 2002). From this perspective, the imagined scene, including the image of the "self," is often perceived as distant in time and/or relevance to the present self (Libby & Eibach, 2002; Nigro & Neisser, 1983; Valenti, Libby, & Eibach, 2011).

While visual perspective clearly influences qualities of our mental representations, there is debate about what determines the visual perspective adopted and, more importantly, what function it serves. A compelling theory proposed by Libby and Eibach (2011) suggests that visual perspective in mental imagery represents the level of construal of the imagined event. From a first-person perspective, significance is given to concrete, experiential aspects of the imagined scenario, whereas from a third-person perspective, significance is given to the overarching personal meaning of the event in relation to its broader context (Libby & Eibach, 2011). To this end, the perspective adopted during mental imagery may reflect the nature of mental state attributions—or whether mental state attributions are occurring at all—with first-person imagery corresponding with more defined, concrete interpretations of targets' mental states (e.g., "He was smiling"), and third-person imagery corresponding with more abstract interpretations (e.g., "He was enjoying the moment") that integrate the motivations, reasons, or context for a target's mental state (Libby, Shaeffer, & Eibach, 2009).

Neural correlates. A large body of evidence finds that processes that engage mental imagery, such as autobiographical memory, prospection, and imagination, exhibit overlapping activity in regions across the DMN, including the dorsal and anterior/ventral mPFC, medial temporal lobe, precuneus, PCC, retrosplenial cortex, TPJ, and superior temporal sulcus (Addis, Wong, & Schacter, 2007; Beaty, Thakral, Madore, Benedek, & Schacter, 2018; Spreng, Mar, & Kim, 2009). However, only a handful of studies have characterized the neural correlates related to adopting different visual perspectives during mental imagery.

Adopting a first-person perspective, whether when recalling episodic memories (Eich, Handy, Holmes, Lerner, & McIsaac, 2012), imagining painful episodes (Christian, Parkinson, Macrae, Miles, & Wheatley, 2015; van der Heiden, Scherpiet, Konicar, Birbaumer, & Veit, 2013), or visualizing action (Ruby & Decety, 2001), shows common activation in the insula and regions of the somatosensory/somatomotor cortex (but see Grol, Vingerhoets, & De Raedt, 2017)—areas implicated in affective salience and interoception (Critchley, Wiens, Rotshtein, Öhman, & Dolan, 2004; Seeley et al., 2007). Across the same paradigms, adopting a third-person perspective is linked to greater activity predominately in the right inferior parietal lobule (IPL) and PCC/precuneus (Grol et al., 2017; Ruby & Decety, 2001; St. Jacques, Szpunar, & Schacter, 2017; van der Heiden et al., 2013; but see Christian et al., 2015; Eich, Nelson, Leghari, & Handy, 2009)—regions of the mentalizing network.

However, it has been suggested that the posterior parietal cortex, particularly the precuneus, may play a key role in *shifting* visual perspectives more generally, rather than adopting a third-person perspective specifically (Ciaramelli, Rosenbaum, Solcz, Levine, & Moscovitch, 2010; St. Jacques et al., 2017; St. Jacques, Carpenter, Szpunar, & Schacter, 2018). Of note, the majority of these studies use tasks in which participants are instructed to recall memories from a certain perspective, which may require more effortful retrieval and result in different neural correlates than observing naturally induced visual perspectives during imagery.

Additional relevant neural evidence comes from studies of emotion regulation that differentiate between “self-immersed” and “self-distanced” perspectives (Kross, Ayduk, & Mischel, 2005), which share similarities with first-person and third-person perspectives, respectively. These constructs have been used to contrast maladaptive versus adaptive methods of reflecting on negative emotions, with self-immersion increasing negative arousal and physiological reactivity and self-distancing reducing it (Kross & Ayduk, 2017; Wang, Yang, Yang, & Huang, 2019). Self-distancing has been found to engage neural regions that overlap with adopting a third-person perspective, including the IPL and PCC/precuneus (Dörfel et al., 2014; Koenigsberg et al., 2010; Ochsner et al., 2004). However, as noted by Libby and Eibach (2011), manipulations intended to promote self-distancing often include instructions to adopt a “detached,” “objective,” or “distant” view, terms which may influence the perceived discrepancy between one’s current and imagined self beyond what would result from spontaneously adopting a third-person perspective. Thus, it is unclear to what degree neural activity related to self-distancing can be generalized to indicate neural activity related to adopting a third-person perspective.

Relevance to social anxiety. Use of the third-person perspective has received particular attention in social anxiety research, given the significant role of negative and distorted self-imagery in the maintenance of SAD (Heimberg et al., 2010; Hirsch, Clark, Mathews, & Williams, 2003; Ng, Abbott, & Hunt, 2014). During social situations, socially anxious individuals form spontaneous images of themselves as a *social object,* imagining from a third-person perspective how others might be seeing them based on their own thoughts, feelings, and internal sensations (Clark & Wells, 1995; Heimberg et al., 2010). Socially anxious people report experiencing these self-focused images not only *during* social situations (Hackmann, Surawy, & Clark, 1998), but also in the period leading up to a social situation (Hinrichsen & Clark, 2003) and in the period after a social situation (D’Argembeau, Van der Linden, d’Acremont, & Mayers, 2006; Ng et al., 2014; Wells, Clark, & Ahmad, 1998). In individuals with SAD, use of the third-person perspective when recalling a social situation becomes even more pronounced over time, whereas non-socially anxious individuals recall social memories predominately from a first-person perspective both immediately after and in the weeks following the event (Coles, Turk, & Heimberg, 2002). Interestingly, when recollecting memories *without* social anxiety-provoking content, individuals with SAD, like their healthy counterparts, engage in mental imagery primarily from the first-person perspective (Heimberg et al., 2010).

The tendency for individuals with SAD to use the third-person perspective during social situations suggests that these situations elicit more abstract processing, including understanding the self in its broader context (Libby & Eibach, 2011). It is possible that individuals with SAD are engaging in more balanced self- and other-focused mentalizing during this mental imagery, such as attempting to understand how one's own behavior might be affecting a social partner. However, because individuals with SAD tend to have more negative self-concepts compared to healthy individuals (Moscovitch, Orr, Rowa, Reimer, & Antony, 2009), they may be more susceptible to detrimental effects from taking a third-person perspective (Libby & Eibach, 2011). Indeed, interview data suggest that social anxiety-related mental imagery in SAD is mainly focused on the self, consisting of negative images of how one might appear to others (e.g., blushing, shaking, looking nervous) (Hackmann et al., 1998; Wild, Hackmann, & Clark, 2008). Thus, it is more likely that excessive use of third-person perspective in SAD reflects abstract processing of negative self-images to support broad, distorted self-focused beliefs—for example, that one is unlikable, an outsider, or a failure.

Use of the third-person perspective may be maladaptive in SAD not only because of the abstract, negative content of this imagery but also because of its use across contexts where it is unhelpful. When healthy individuals, as well as socially anxious individuals, are instructed to adopt a third-person perspective during a speech performance, they report increased negative thoughts and poorer self-evaluations of their performance (Spurr & Stopa, 2003). This suggests that, for anyone, adopting a third-person perspective during an anxiety-provoking, performance-based situation—as individuals with SAD often do (Hackmann et al., 1998)—may be disruptive, as it indicates attempts to assess broader abstract meaning during a situation in which more concrete, experiential processing may be advantageous.

Integrating Constructs in the Real World

In this chapter, we have delineated three constructs to inform our understanding of the role of the self in mentalizing: the *target* of mentalizing, the *source* of mentalizing representations, and the *visual perspective* used in mental imagery (summarized in Table 1). These constructs share similarities in terms of their neural and psychological correlates—for example, processing aspects relevant to the self tends to be less cognitively demanding than processing aspects relevant to others, and a greater degree of self-processing in one construct likely correlates with a greater degree of self-processing in other constructs. Although similar, each construct describes a unique aspect of mentalizing, and it is likely that these constructs must flexibly work together to facilitate adaptive social cognition. We hypothesize that within each of these constructs, people shift between emphasizing the *self* or the *other* in a dynamic fashion that is influenced by external factors (e.g., the context, topic of conversation, and people involved), by internal factors (e.g., one's mood, physical state, and beliefs), and by these constructs' interactive effects on each

Table 1 Correlates of *self*-focus in constructs related to self-other processing in mentalizing

Construct	Definition	Similar terms	Key brain regions involved	Relevant clinical disorders
Self as target vs. other as target	Inferring one’s own mental states	Self-focus, self-reflection, self-referential thought, introspection	↑ amPFC/vmPFC, dmPFC, ACC, PCC	SAD, MDD, GAD, PTSD
Self as source vs. other as source	Using one’s own mental state as basis for inference about another’s	Egocentricity bias, failure to inhibit self-perspective, low self-other distinction, low self-other control, high self-other overlap, self-projection	↓ IFG, dlPFC, rTPJ/rSMG	SAD, MDD, GAD, ASD, SZ, PD
Self as object vs. self as subject	Viewing oneself in mental imagery, as if from an observer’s perspective	Third-person perspective, observer perspective, self-distanced perspective	↑ rIPL, PCC, precuneus	SAD, MDD, PTSD, BDD, SZ

Note. Due to space limitations, relevant citations can be found in the text
Abbreviations: amPFC anterior medial prefrontal cortex, *vmPFC* ventromedial prefrontal cortex, *dmPFC* dorsomedial prefrontal cortex, *ACC* anterior cingulate cortex, *PCC* posterior cingulate cortex, *IFG* inferior frontal gyrus, *dlPFC* dorsolateral prefrontal cortex, *rTPJ* right temporoparietal junction, *rSMG* right supramarginal gyrus, *rIPL* right inferior parietal lobule, *SAD* social anxiety disorder, *MDD* major depressive disorder, *GAD* generalized anxiety disorder, *PTSD* post-traumatic stress disorder, *ASD* autism spectrum disorder, *SZ* schizophrenia, *PD* personality disorders, *BDD* body dysmorphic disorder

other. Psychological disorders, like SAD, may be related to difficulty flexibly shifting between *self* and *other* within these constructs when it is contextually appropriate. To lend clarity to how these constructs might unfold and interact in a clinically healthy individual, consider the following scenario, also illustrated in Fig. 1a:

Helen is at the grocery store and runs into Steve, an old friend from college. Helen hasn’t seen Steve for several months and asks how he is doing. From his flat expression and vague reply, Helen can tell he is not doing well [other as target]. She reflects on her discomfort [self as target] about potentially probing into his personal life in the middle of the grocery store and chooses to stick with lighter content for now. The conversation turns to reminiscing about the last time they saw each other—at Helen’s former college roommate’s wedding last year—and Steve mentions the memorable toast Helen gave to the new bride and groom. Helen begins to recall her experience during this event [self as target], seeing, as if through her own eyes, the sea of guests as she clutches a glass of champagne and begins her toast [self as subject]. She relives the initial twinge of nervousness and subsequent delight as she visualizes the audience roaring with laughter at tales of her and her college roommate’s youthful shenanigans. While basking in the glow of this memory [self as target], Helen notices a sad smile on Steve’s face [other as target] and realizes he may have had a different experience that night [other as source]. Concerned, she asks if he is okay. His smile vanishes and his eyes well up as he reveals that at the

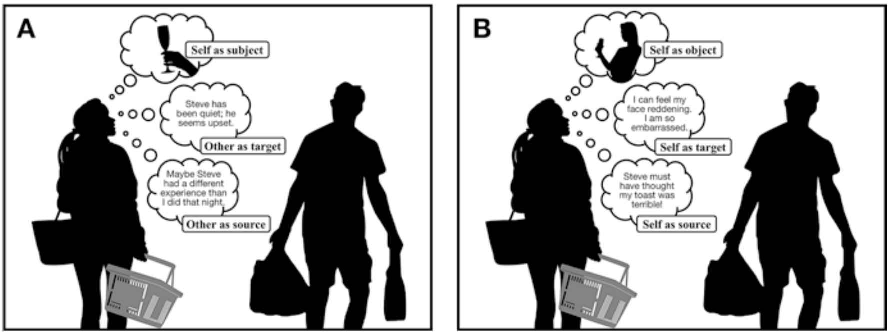

A
Self as subject
Steve has been quiet; he seems upset.
Other as target
Maybe Steve had a different experience than I did that night.
Other as source
B
Self as object
I can feel my face reddening. I am so embarrassed.
Self as target
Steve must have thought my toast was terrible!
Self as source

Fig. 1 Constructs related to self-other processing in mentalizing as they might unfold during a social interaction in (**a**) a clinically healthy individual and (**b**) a socially anxious individual. In (**a**), the healthy individual engages in a healthy balance of self- and other-related mentalizing processes while interacting with her friend, whereas in (**b**), the socially anxious individual overrelies on self-focused mentalizing processes to the ultimate detriment of her social interaction. See text for detailed vignettes

wedding, he discovered his now ex-partner in the coatroom with the caterer. Helen remembers her own pain following a recent messy breakup and imagines Steve must be feeling similarly [self as source]. Helen takes Steve's hand and leads him to the beer and wine aisle.

This relatively brief but complex scene exhibits the ongoing dynamics of adaptively shifting between self and other in relation to the target, source, and visual perspective adopted during mentalizing. In this scene, Helen flexibly switches between reflecting on her own mental states and the mental states of her friend Steve. She recalls the memory of her wedding toast from an embodied first-person perspective, allowing her to experientially relive the emotions of that night. Even during this memory, she notices Steve's sad expression and is able to inhibit her own perspective to infer his differing emotional state. Later, Helen draws from her own similar experience and makes an egocentric inference about his mental state following a breakup, allowing her to understand his experience with minimal effort.

Now, consider the scenario from the perspective of a socially anxious individual (Fig. 1b):

Helen is at the grocery store and runs into Steve, an old friend from college. Helen hasn't seen Steve for several months and starts to feel anxious, so she monitors how she feels [self as target] and conjures an image from an outsider's perspective of how she might look [self as object] to make sure she doesn't embarrass herself. Helen asks how Steve is doing, and drawing from her own feelings of discomfort, she can tell he is unhappy to see her [self as source]. The conversation turns to reminiscing about the last time they saw each other—at Helen's former college roommate's wedding last year—and Steve mentions the memorable toast Helen gave to the new bride and groom. Helen begins to recall this event as if a member of the audience, watching herself stand in front of a sea of guests with her hand trembling as she clutches her glass of champagne [self as object]. She *remembers people laughing during her toast but suspects it was either out of surprise at realizing she has a sense of humor, or—worse yet—pity [self as source]. Helen mutters, “Yeah, it was a pretty terrible toast.” Steve looks confused by this statement and replies, “Are you kidding? It was great!” He frowns and adds, “Definitely better than finding my partner in the coatroom with the caterer...” Helen interprets his response as a jab at her toast [self as source] and feels the heat in her face as she experiences intense embarrassment [self as target]. She quickly makes an excuse to leave the conversation and heads toward the checkout line. Steve looks on, bewildered by Helen’s abrupt exit following his attempt at disclosing his painful breakup.*

In this scenario, Helen appears to over-rely on self-related mentalizing processes. She focuses primarily on examining her own mental states rather than inferring Steve’s; she projects her own thoughts and feelings onto Steve instead of working to understand his possibly differing view; she elicits images of herself as a social object when recalling past experiences. As a result, she misses important cues from Steve, egocentrically misinterprets his mental states, and relies on distorted, abstract mental images of herself to guide her behavior. Ultimately, this self-focus will likely prevent Helen from finding evidence to disconfirm her negative self-image, serving to perpetuate her social anxiety in future interactions.

Broader Clinical Implications

We have focused on the pathology of SAD due to its strong empirical evidence of dysfunctional self- and other-processing in mentalizing. However, each of the three self/other constructs described in this chapter has been linked to several other psychological disorders. Although a detailed discussion of the role of these constructs across psychological disorders is beyond the scope of this chapter, we will briefly touch on particularly relevant disorders here, which are also highlighted in Table 1. Heightened focus on the self as a *target* of mentalizing, as seen in ruminative self-focus (Moberly & Watkins, 2008; Treynor, Gonzalez, & Nolen-Hoeksema, 2003), is widely recognized as a transdiagnostic marker (Andrews-Hanna et al., 2020; Kaplan et al., 2018) and features prominently in disorders including depression (Watkins & Teasdale, 2004), anxiety symptoms (McLaughlin & Nolen-Hoeksema, 2011), and post-traumatic stress disorder (PTSD; Michael, Halligan, Clark, & Ehlers, 2007). Overreliance on the self as the *source* of mentalizing representations, resulting in egocentricity biases, is seen in depression (Erle, Barth, & Topolinski, 2018; Hoffmann, Banzhaf, et al., 2016), anxiety symptoms (Todd et al., 2015), autism spectrum disorders (ASD; Hoffmann, Koehne, et al., 2016), psychopathy (Bresin, Boyd, Ode, & Robinson, 2013), and schizophrenia (van der Weiden, Prikken, & van Haren, 2015). A greater tendency toward adopting a third-person, *other* perspective during mental imagery has been identified in depression (Lemogne et al., 2006), PTSD (Berntsen, Willert, & Rubin, 2003), body dysmorphic disorder (Osman, Cooper, Hackman, & Veale, 2004), schizophrenia (Potheegadoo, Berna, Cuervo-Lombard, & Danion, 2013), and narcissistic personality disorder (Marchlewska & Cichocka, 2017). Thus, overreliance on the self as the target, source, or object of visual perspective may be shared features across many psychological disorders, warranting research on their utility as transdiagnostic markers.

Given their presence across multiple disorders, these constructs may provide useful targets for therapeutic intervention. Indeed, some empirically supported treatments have already been found to alter these constructs. For example, mindfulness-based therapies may reduce maladaptive self-focused mentalizing (Baer, 2009); mentalization-based therapy may reduce egocentricity biases resulting from poor self-other differentiation (Fonagy & Luyten, 2009); and imagery rescripting techniques may rely on adaptive use of visual perspectives to change the meaning of negative self-related imagery (Çili & Stopa, 2015; Lee & Kwon, 2013). In future work, it will be important to identify whether heightened self-focus in each construct serves as a cause or correlate of dysfunction in the psychological disorders in which they are seen.

Future Directions

We have attempted to integrate research on the many ways emphasis on the *self* can impact mental state inferences about *others*, with the goal of improving our awareness of what we know—and don't know—about functional and dysfunctional mentalizing. As is evident, many avenues remain to be explored. To date, most research on self- and other-focused mentalizing involves tightly controlled, laboratory-based studies, providing little real-world understanding of how mentalizing processes naturally occur across different social and non-social contexts, including whether and how different categories of "others"—such as strangers, acquaintances, friends, or partners—correspond with alterations in mentalizing. Additionally, little is known about how these processes and their neural underpinnings unfold *dynamically*, either in the short term or long term, or how they change developmentally within individuals.

Although dysfunctions in constructs related to self- and other-processing are linked to multiple psychological disorders, it is unclear whether they are causal factors in initiating and/or maintaining mental illness or simply correlates of mental illness, warranting careful research into the possible mechanistic role of mentalizing deficits in psychopathology. Relatedly, future work may want to examine what additional factors interact with dysfunctional self/other processing to yield divergent psychological disorders. For example, both social anxiety and depression are related to a greater tendency toward adopting a third-person perspective; however, individuals high in social anxiety may be more likely to adopt this perspective when imagining social situations (D'Argembeau et al., 2006), whereas those high in depression may be more likely to adopt the perspective when recalling positive autobiographical events (Lemogne et al., 2006; Nelis, Debeer, Holmes, & Raes, 2013). Finally, despite substantial research examining alterations in self- and other-processing in psychopathology, a dearth of studies links this research to the brain. Broadening our understanding of the neural correlates of self-other processes in clinical populations would both aid in identifying targets for treatment across multiple disorders and contribute to our understanding of adaptive and maladaptive mentalizing.

In sum, the interplay between self and other processes and their impact on mentalizing can be illustrated through multiple constructs. While we have detailed three such constructs here—the *target* of mentalizing, the *source* of mentalizing, and the *visual perspective* adopted during mentalizing-related mental imagery—there are likely many other ways of conceptualizing self/other differences during mental state inference. These constructs share similarities in terms of neural and psychological correlates but also provide unique contributions to mental state attribution, working together dynamically to produce adaptive mentalizing. Dysfunctions in these constructs, such as overreliance on the self, may contribute to mentalizing deficits and other symptoms seen across psychological disorders, including SAD. Identifying and expanding on the precise ways that processes related to the self and other differ, overlap, and interact will likely be necessary to attain a complete understanding of mentalizing, including its basic mechanisms, the ways in which it can go awry, and how it can be treated effectively.

References

Addis, D. R., Wong, A. T., & Schacter, D. L. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia, 45*, 1363. https://doi.org/10.1016/j.neuropsychologia.2006.10.016

Alden, L. E., Auyeung, K. W., & Plasencia, L. (2014). Social anxiety and the self. In S. G. Hofmann & P. M. DiBartolo (Eds.), *Social anxiety: Clinical, developmental, and social perspectives* (3rd ed., pp. 531–549). London, England: Academic Press. https://doi.org/10.1016/B978-0-12-394427-6.00018-2

Alden, L. E., Regambal, M. J., & Plasencia, L. (2014). Relational processes in social anxiety disorder. In J. W. Weeks (Ed.), *The Wiley Blackwell handbook of social anxiety disorder*. Chichester, England: Wiley. https://doi.org/10.1002/9781118653920.ch8

Ames, D. R. (2004). Strategies for social inference: A similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of Personality and Social Psychology, 87*, 573. https://doi.org/10.1037/0022-3514.87.5.573

Ames, D. R., Weber, E. U., & Zou, X. (2012). Mind-reading in strategic interaction: The impact of perceived similarity on projection and stereotyping. *Organizational Behavior and Human Decision Processes, 117*, 96. https://doi.org/10.1016/j.obhdp.2011.07.007

Andrews-Hanna, J. R., Christoff, K., & O'Connor, M. (2020). Dynamic regulation of internal experience: Mechanisms of therapeutic change. In R.D. Lane & L. Nadel (Eds.), *The neuroscience of enduring change: Implications for psychotherapy* (pp. 89-131). Oxford University Press.

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., & Buckner, R. L. (2010). Functional-anatomic fractionation of the brain's default network. *Neuron, 65*, 550. https://doi.org/10.1016/j.neuron.2010.02.005

Andrews-Hanna, J. R., Smallwood, J., & Spreng, R. N. (2014). The default network and self-generated thought: Component processes, dynamic control, and clinical relevance. *Annals of the New York Academy of Sciences, 1316*(1), 29–52. https://doi.org/10.1111/nyas.12360

Apperly, I. A., Back, E., Samson, D., & France, L. (2008). The cost of thinking about false beliefs: Evidence from adults' performance on a non-inferential theory of mind task. *Cognition, 106*, 1093. https://doi.org/10.1016/j.cognition.2007.05.005

Aron, A., Lewandowski, G. W., Mashek, D., & Aron, E. N. (2013). *The self-expansion model of motivation and cognition in close relationships*. Oxford, England: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195398694.013.0005

Auyeung, K. W., & Alden, L. E. (2016). Social anxiety and empathy for social pain. *Cognitive Therapy and Research, 40*, 38. https://doi.org/10.1007/s10608-015-9718-0

Baer, R. A. (2009). Self-focused attention and mechanisms of change in mindfulness-based treatment. *Cognitive Behaviour Therapy, 38*, 15. https://doi.org/10.1080/16506070902980703

Baetens, K., Ma, N., Steen, J., & Van Overwalle, F. (2013). Involvement of the mentalizing network in social and non-social high construal. *Social Cognitive and Affective Neuroscience, 9*, 817. https://doi.org/10.1093/scan/nst048

Baetens, K., Ma, N., & Van Overwalle, F. (2017). The dorsal medial prefrontal cortex is recruited by high construal of non-social stimuli. *Frontiers in Behavioral Neuroscience, 11*, 44. https://doi.org/10.3389/fnbeh.2017.00044

Bar, M. (2009). The proactive brain: Memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*, 1235. https://doi.org/10.1098/rstb.2008.0310

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage, 76*, 412. https://doi.org/10.1016/j.neuroimage.2013.02.063

Beaty, R. E., Thakral, P. P., Madore, K. P., Benedek, M., & Schacter, D. L. (2018). Core network contributions to remembering the past, imagining the future, and thinking creatively. *Journal of Cognitive Neuroscience, 30*, 1939. https://doi.org/10.1162/jocn_a_01327

Beitel, M., Ferrer, E., & Cecero, J. J. (2005). Psychological mindedness and awareness of self and others. *Journal of Clinical Psychology, 61*, 739. https://doi.org/10.1002/jclp.20095

Bem, D. J. (1972). Self-perception theory. *Advances in Experimental Social Psychology, 67*, 1–62. https://doi.org/10.1016/S0065-2601(08)60024-6

Berntsen, D., Willert, M., & Rubin, D. C. (2003). Splintered memories or vivid landmarks? Qualities and organization of traumatic memories with and without PTSD. *Applied Cognitive Psychology., 17*, 675. https://doi.org/10.1002/acp.894

Blair, K. S., Geraci, M., Devido, J., McCaffrey, D., Chen, G., Vythilingam, M., … Pine, D. S. (2008). Neural response to self- and other referential praise and criticism in generalized social phobia. *Archives of General Psychiatry, 65*(10), 1176–1184. https://doi.org/10.1001/archpsyc.65.10.1176

Blair, K. S., Geraci, M., Otero, M., Majestic, C., Odenheimer, S., Jacobs, M., … Pine, D. S. (2011). Atypical modulation of medial prefrontal cortex to self-referential comments in generalized social phobia. *Psychiatry Research - Neuroimaging, 193*, 38. https://doi.org/10.1016/j.pscychresns.2010.12.016

Boehme, S., Ritter, V., Tefikow, S., Stangier, U., Strauss, B., Miltner, W. H. R., & Straube, T. (2015). Neural correlates of emotional interference in social anxiety disorder. *PLoS One, 10*, e0128608. https://doi.org/10.1371/journal.pone.0128608

Bollich, K. L., Johannet, P. M., & Vazire, S. (2011). In search of our true selves: Feedback as a path to self-knowledge. *Frontiers in Psychology, 2*, 312. https://doi.org/10.3389/fpsyg.2011.00312

Bradford, E. E. F., Jentzsch, I., & Gomez, J. C. (2015). From self to social cognition: Theory of mind mechanisms and their relation to executive functioning. *Cognition, 138*, 21. https://doi.org/10.1016/j.cognition.2015.02.001

Braga, R. M., & Buckner, R. L. (2017). Parallel interdigitated distributed networks within the individual estimated by intrinsic functional connectivity. *Neuron, 95*, 457. https://doi.org/10.1016/j.neuron.2017.06.038

Bresin, K., Boyd, R. L., Ode, S., & Robinson, M. D. (2013). Egocentric perceptions of the environment in primary, but not secondary, psychopathy. *Cognitive Therapy and Research, 37*, 412. https://doi.org/10.1007/s10608-012-9459-2

Brühl, A. B., Rufer, M., Delsignore, A., Kaffenberger, T., Jäncke, L., & Herwig, U. (2011). Neural correlates of altered general emotion processing in social anxiety disorder. *Brain Research, 1378*, 72. https://doi.org/10.1016/j.brainres.2010.12.084

Bryant, L., Coffey, A., Povinelli, D. J., & Pruett, J. R. (2013). Theory of mind experience sampling in typical adults. *Consciousness and Cognition, 22*, 697. https://doi.org/10.1016/j.concog.2013.04.005

Buhlmann, U., Wacker, R., & Dziobek, I. (2015). Inferring other people's states of mind: Comparison across social anxiety, body dysmorphic, and obsessive-compulsive disorders. *Journal of Anxiety Disorders, 34*, 107. https://doi.org/10.1016/j.janxdis.2015.06.003

Butler, A. C., Rice, H. J., Wooldridge, C. L., & Rubin, D. C. (2016). Visual imagery in autobiographical memory: The role of repeated retrieval in shifting perspective. *Consciousness and Cognition, 42*, 237. https://doi.org/10.1016/j.concog.2016.03.018

Chavez, R. S., Heatherton, T. F., & Wagner, D. D. (2017). Neural population decoding reveals the intrinsic positivity of the self. *Cerebral Cortex, 27*(11), 5222–5229. https://doi.org/10.1093/cercor/bhw302

Christian, B. M., Parkinson, C., Macrae, C. N., Miles, L. K., & Wheatley, T. (2015). When imagining yourself in pain, visual perspective matters: The neural and behavioral correlates of simulated sensory experiences. *Journal of Cognitive Neuroscience, 27*, 866. https://doi.org/10.1162/jocn_a_00754

Ciaramelli, E., Rosenbaum, R. S., Solcz, S., Levine, B., & Moscovitch, M. (2010). Mental space travel: Damage to posterior parietal cortex prevents egocentric navigation and reexperiencing of remote spatial memories. *Journal of Experimental Psychology: Learning Memory and Cognition, 36*, 619. https://doi.org/10.1037/a0019181

Çili, S., & Stopa, L. (2015). Intrusive mental imagery in psychological disorders: Is the self the key to understanding maintenance? *Frontiers in Psychiatry, 6*, 103. https://doi.org/10.3389/fpsyt.2015.00103

Clark, D. M., & Wells, A. (1995). A cognitive model of social phobia. *Social Phobia: Diagnosis, Assessment, and Treatment, 41*, 68.

Coan, J. A., & Maresh, E. L. (2014). Social baseline theory and the social regulation of emotion. In J. J. Gross (Ed.), *Handbook of emotion regulation* (2nd ed.). New York, NY: Guilford Press.

Coles, M. E., Turk, C. L., & Heimberg, R. G. (2002). The role of memory perspective in social phobia: Immediate and delayed memories for role-played situations. *Behavioural and Cognitive Psychotherapy, 30*, 415. https://doi.org/10.1017/S1352465802004034

Cooley, C. H. (1909). *Two major works: Social organization. Human nature and the social order*. Glencoe, IL: Free Press.

Cotter, J., Granger, K., Backx, R., Hobbs, M., Looi, C. Y., & Barnett, J. H. (2018). Social cognitive dysfunction as a clinical marker: A systematic review of meta-analyses across 30 clinical conditions. *Neuroscience and Biobehavioral Reviews, 84*, 92. https://doi.org/10.1016/j.neubiorev.2017.11.014

Critchley, H. D., Wiens, S., Rotshtein, P., Öhman, A., & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience, 7*(2), 189–195. https://doi.org/10.1038/nn1176

D'Argembeau, A. (2013). On the role of the ventromedial prefrontal cortex in self-processing: The valuation hypothesis. *Frontiers in Human Neuroscience, 7*, 372. https://doi.org/10.3389/fnhum.2013.00372

D'Argembeau, A., Van der Linden, M., d'Acremont, M., & Mayers, I. (2006). Phenomenal characteristics of autobiographical memories for social and non-social events in social phobia. *Memory, 14*, 637. https://doi.org/10.1080/09658210600747183

Damasio, A. (2010). *Self comes to mind: Constructing the conscious brain*. New York, NY: Pantheon Books.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*, 113. https://doi.org/10.1037/0022-3514.44.1.113

Davis, M. H., & Franzoi, S. L. (1991). Stability and change in adolescent self-consciousness and empathy. *Journal of Research in Personality, 25*, 70. https://doi.org/10.1016/0092-6566(91)90006-C

de Guzman, M., Bird, G., Banissy, M. J., & Catmur, C. (2016). Self-other control processes in social cognition: From imitation to empathy. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*, 20150079. https://doi.org/10.1098/rstb.2015.0079

De Vignemont, F., & Mercier, H. (2016). Under influence: Is altercentric bias compatible with simulation theory? In H. Kornblith & B. McLaughlin (Eds.), *Alvin Goldman and his critics*. Oxford, England: Blackwell. https://doi.org/10.1002/9781118609378.ch13

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience, 24*, 1742. https://doi.org/10.1162/jocn_a_00233

Dimaggio, G., Lysaker, P. H., Carcione, A., Nicolò, G., & Semerari, A. (2008). Know yourself and you shall know the other... to a certain extent: Multiple paths of influence of self-reflection on mindreading. *Consciousness and Cognition, 17*, 778. https://doi.org/10.1016/j.concog.2008.02.005

Dörfel, D., Lamke, J. P., Hummel, F., Wagner, U., Erk, S., & Walter, H. (2014). Common and differential neural networks of emotion regulation by detachment, reinterpretation, distraction, and expressive suppression: A comparative fMRI investigation. *NeuroImage, 101*, 298. https://doi.org/10.1016/j.neuroimage.2014.06.051

Eich, E., Handy, T. C., Holmes, E. A., Lerner, J., & McIsaac, H. K. (2012). Field and observer perspectives in autobiographical memory. In J. P. Forgas, K. Fiedler, & C. Sedikides (Eds.), *Social thinking and interpersonal behavior* (pp. 163–181). New York, NY: Taylor and Francis. https://doi.org/10.4324/9780203139677

Eich, E., Nelson, A. L., Leghari, M. A., & Handy, T. C. (2009). Neural systems mediating field and observer memories. *Neuropsychologia, 47*, 2239. https://doi.org/10.1016/j.neuropsychologia.2009.02.019

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science, 17*, 311. https://doi.org/10.1111/j.1467-9280.2006.01704.x

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87*, 327. https://doi.org/10.1037/0022-3514.87.3.327

Erle, T. M., Barth, N., & Topolinski, S. (2018). Egocentrism in sub-clinical depression. *Cognition and Emotion, 33*(6), 1239–1248.

Eyal, T., Steffel, M., & Epley, N. (2018). Perspective mistaking: Accurately understanding the mind of another requires getting perspective, not taking perspective. *Journal of Personality and Social Psychology, 114*, 547. https://doi.org/10.1037/pspa0000115

Eysenck, M. W., & Derakshan, N. (2011). New perspectives in attentional control theory. *Personality and Individual Differences, 50*(7), 955–960. https://doi.org/10.1016/j.paid.2010.08.019

Fenigstein, A., Scheier, M. F., & Buss, A. H. (1975). Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology, 43*, 522. https://doi.org/10.1037/h0076760

Fonagy, P., Gergely, G., Jurist, E. L., & Target, M. (2002). *Affect regulation, mentalization, and the development of the self*. New York, NY: Other Press. Retrieved from https://psycnet.apa.org/record/2002-17653-000

Fonagy, P., & Luyten, P. (2009). A developmental, mentalization-based approach to the understanding and treatment of borderline personality disorder. *Development and Psychopathology, 21*, 1355. https://doi.org/10.1017/S0954579409990198

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50*, 531. https://doi.org/10.1016/j.neuron.2006.05.001

Galinsky, A. D., Ku, G., & Wang, C. S. (2005). Perspective-taking and self-other overlap: Fostering social bonds and facilitating social coordination. *Group Processes and Intergroup Relations, 8*, 109. https://doi.org/10.1177/1368430205051060

Garvert, M. M., Moutoussis, M., Kurth-Nelson, Z., Behrens, T. E. J., & Dolan, R. J. (2015). Learning-induced plasticity in medial prefrontal cortex predicts preference malleability. *Neuron, 85*, 418. https://doi.org/10.1016/j.neuron.2014.12.033

Gerace, A., Day, A., Casey, S., & Mohr, P. (2017). 'I think you think': Understanding the importance of self-reflection to the taking of another person's perspective. *Journal of Relationships Research, 8*, e9. https://doi.org/10.1017/jrr.2017.8

Gilead, M., Trope, Y., & Liberman, N. (2019). Above and beyond the concrete: The diverse representational substrates of the predictive brain. *Behavioral and Brain Sciences, 43*, 1–63.

Gordon, R. M. (1986). Folk psychology as simulation. *Mind & Language., 1*, 158–171. https://doi.org/10.1111/j.1468-0017.1986.tb00324.x

Grol, M., Vingerhoets, G., & De Raedt, R. (2017). Mental imagery of positive and neutral memories: A fMRI study comparing field perspective imagery to observer perspective imagery. *Brain and Cognition, 111*, 13. https://doi.org/10.1016/j.bandc.2016.09.014

Gros, D. F., & Sarver, N. W. (2014). An investigation of the psychometric properties of the Social Thoughts and Beliefs Scale (STABS) and structure of cognitive symptoms in participants with social anxiety disorder and healthy controls. *Journal of Anxiety Disorders, 28*, 283. https://doi.org/10.1016/j.janxdis.2014.01.004

Hackmann, A., Surawy, C., & Clark, D. M. (1998). Seeing yourself through others' eyes: A study of spontaneously occurring images in social phobia. *Behavioural and Cognitive Psychotherapy, 26*, 3. https://doi.org/10.1017/S1352465898000022

Hartwright, C. E., Apperly, I. A., & Hansen, P. C. (2012). Multiple roles for executive control in belief-desire reasoning: Distinct neural networks are recruited for self perspective inhibition and complexity of reasoning. *NeuroImage, 61*, 921. https://doi.org/10.1016/j.neuroimage.2012.03.012

Heimberg, R. G., Brozovich, F. A., & Rapee, R. M. (2010). A cognitive behavioral model of social anxiety disorder: Update and extension. In S. G. Hofmann & P. M. DiBartolo (Eds.), *Social anxiety: Clinical, developmental, and social perspectives* (pp. 395–422). New York, NY: Academic Press. https://doi.org/10.1016/B978-0-12-375096-9.00015-8

Heitmann, C. Y., Feldker, K., Neumeister, P., Brinkmann, L., Schrammen, E., Zwitserlood, P., & Straube, T. (2017). Brain activation to task-irrelevant disorder-related threat in social anxiety disorder: The impact of symptom severity. *NeuroImage: Clinical, 14*, 323. https://doi.org/10.1016/j.nicl.2017.01.020

Heitmann, C. Y., Feldker, K., Neumeister, P., Zepp, B. M., Peterburs, J., Zwitserlood, P., & Straube, T. (2016). Abnormal brain activation and connectivity to standardized disorder-related visual scenes in social anxiety disorder. *Human Brain Mapping, 37*, 1559. https://doi.org/10.1002/hbm.23120

Hezel, D. M., & McNally, R. J. (2014). Theory of mind impairments in social anxiety disorder. *Behavior Therapy, 45*, 530. https://doi.org/10.1016/j.beth.2014.02.010

Hinrichsen, H., & Clark, D. M. (2003). Anticipatory processing in social anxiety: Two pilot studies. *Journal of Behavior Therapy and Experimental Psychiatry, 34*, 205. https://doi.org/10.1016/S0005-7916(03)00050-8

Hirsch, C. R., Clark, D. M., Mathews, A., & Williams, R. (2003). Self-images play a causal role in social phobia. *Behaviour Research and Therapy, 41*, 909. https://doi.org/10.1016/S0005-7967(02)00103-1

Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology, 53*, 221. https://doi.org/10.1037/0022-3514.53.2.221

Hoffmann, F., Banzhaf, C., Kanske, P., Gärtner, M., Bermpohl, F., & Singer, T. (2016). Empathy in depression: Egocentric and altercentric biases and the role of alexithymia. *Journal of Affective Disorders, 199*, 23. https://doi.org/10.1016/j.jad.2016.03.007

Hoffmann, F., Koehne, S., Steinbeis, N., Dziobek, I., & Singer, T. (2016). Preserved self-other distinction during empathy in autism is linked to network integrity of right supramarginal gyrus. *Journal of Autism and Developmental Disorders, 46*, 637. https://doi.org/10.1007/s10803-015-2609-0

Holmes, E. A., Coughtrey, A. E., & Connor, A. (2008). Looking at or through rose-tinted glasses? *Imagery Perspective and Positive Mood. Emotion, 8*, 875. https://doi.org/10.1037/a0013617

Honeycutt, J. M., & Ford, S. G. (2006). Chapter 9: Mental imagery and intrapersonal communication: A review of research on imagined interactions (IIs) and current developments. *Communication Yearbook, 25*, 315. https://doi.org/10.1207/s15567419cy2501_9

Hope, D. A., Burns, J. A., Hayes, S. A., Herbert, J. D., & Warner, M. D. (2010). Automatic thoughts and cognitive restructuring in cognitive behavioral group therapy for social anxiety disorder. *Cognitive Therapy and Research, 34*, 1. https://doi.org/10.1007/s10608-007-9147-9

Igelstrom, K. M., Webb, T. W., & Graziano, M. S. A. (2015). Neural processes in the human temporoparietal cortex separated by localized independent component analysis. *Journal of Neuroscience, 35*, 9432. https://doi.org/10.1523/jneurosci.0551-15.2015

James, W. (1890). *The principles of psychology* (Vol. 1 & 2). New York, NY: Holt. https://doi.org/10.1037/10538-000

Kaplan, D. M., Palitsky, R., Carey, A. L., Crane, T. E., Havens, C. M., Medrano, M. R., … OConnor, M. F. (2018). Maladaptive repetitive thought as a transdiagnostic phenomenon and treatment target: An integrative review. *Journal of Clinical Psychology, 74*, 1126. https://doi.org/10.1002/jclp.22585

Kappesser, J., Williams, A. C., & Prkachin, K. M. (2006). Testing two accounts of pain underestimation. *Pain, 124*, 109. https://doi.org/10.1016/j.pain.2006.04.003

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*, 32. https://doi.org/10.1111/1467-9280.00211

Koenigsberg, H. W., Fan, J., Ochsner, K. N., Liu, X., Guise, K., Pizzarello, S., … Siever, L. J. (2010). Neural correlates of using distancing to regulate emotional responses to social situations. *Neuropsychologia, 48*, 1813. https://doi.org/10.1016/j.neuropsychologia.2010.03.002

Koster-Hale, J., Richardson, H., Velez, N., Asaba, M., Young, L., & Saxe, R. (2017). Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. *NeuroImage, 161*, 9. https://doi.org/10.1016/j.neuroimage.2017.08.026

Krienen, F. M., Tu, P.-C., & Buckner, R. L. (2010). Clan mentality: Evidence that the medial prefrontal cortex responds to close others. *Journal of Neuroscience, 30*, 13906. https://doi.org/10.1523/jneurosci.2180-10.2010

Kross, E., & Ayduk, O. (2017). Self-distancing: Theory, research, and current directions. In J. M. Olson (Ed.), *Advances in experimental social psychology* (Vol. 55, pp. 81–136). New York, NY: Academic Press. https://doi.org/10.1016/bs.aesp.2016.10.002

Kross, E., Ayduk, O., & Mischel, W. (2005). When asking "why" does not hurt: Distinguishing rumination from reflective processing of negative emotions. *Psychological Science, 16*, 709. https://doi.org/10.1111/j.1467-9280.2005.01600.x

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology, 67*, 596. https://doi.org/10.1037/0022-3514.67.4.596

Kumaran, D., Banino, A., Blundell, C., Hassabis, D., & Dayan, P. (2016). Computations underlying social hierarchy learning: Distinct neural mechanisms for updating and representing self-relevant information. *Neuron, 92*, 1135. https://doi.org/10.1016/j.neuron.2016.10.052

Lamm, C., Bukowski, H., & Silani, G. (2016). From shared to distinct self-other representations in empathy: Evidence from neurotypical function and socio-cognitive disorders. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1686), 20150083. https://doi.org/10.1098/rstb.2015.0083

Lamm, C., & Majdandžić, J. (2015). The role of shared neural activations, mirror neurons, and morality in empathy—A critical comment. *Neuroscience Research, 90*, 15. https://doi.org/10.1016/j.neures.2014.10.008

Lee, S. W., & Kwon, J. H. (2013). The efficacy of imagery rescripting (IR) for social phobia: A randomized controlled trial. *Journal of Behavior Therapy and Experimental Psychiatry, 44*, 351. https://doi.org/10.1016/j.jbtep.2013.03.001

Legrand, D., & Ruby, P. (2009). What is self-specific? Theoretical investigation and critical review of neuroimaging results. *Psychological Review, 116*, 252. https://doi.org/10.1037/a0014172

Lemogne, C., Piolino, P., Friszer, S., Claret, A., Girault, N., Jouvent, R., … Fossati, P. (2006). Episodic autobiographical memory in depression: Specificity, autonoetic consciousness, and self-perspective. *Consciousness and Cognition, 15*, 258. https://doi.org/10.1016/j.concog.2005.07.005

Lenton-Brym, A. P., Moscovitch, D. A., Vidovic, V., Nilsen, E., & Friedman, O. (2018). Theory of mind ability in high socially anxious individuals. *Anxiety, Stress and Coping, 31*, 487. https://doi.org/10.1080/10615806.2018.1483021

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in "theory of mind". *Trends in Cognitive Sciences, 8*, 528. https://doi.org/10.1016/j.tics.2004.10.001

Libby, L. K., & Eibach, R. P. (2002). Looking back in time: Self-concept change affects visual perspective in autobiographical memory. *Journal of Personality and Social Psychology, 82*, 167. https://doi.org/10.1037/0022-3514.82.2.167

Libby, L. K., & Eibach, R. P. (2011). Visual perspective in mental imagery. A representational tool that functions in judgment, emotion, and self-insight. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 44, pp. 185–245). New York, NY: Academic Press. https://doi.org/10.1016/B978-0-12-385522-0.00004-4

Libby, L. K., & Eibach, R. P. (2013). The role of visual imagery in social cognition. In D. E. Carlston (Ed.), *Oxford library of psychology. The Oxford handbook of social cognition* (pp. 147–166). Oxford, England: Oxford University Press.

Libby, L. K., Shaeffer, E. M., & Eibach, R. P. (2009). Seeing meaning in action: A bidirectional link between visual perspective and action identification level. *Journal of Experimental Psychology: General, 138*, 503. https://doi.org/10.1037/a0016795

Libby, L. K., Valenti, G., Pfent, A., & Eibach, R. P. (2011). Seeing failure in your life: Imagery perspective determines whether self-esteem shapes reactions to recalled and imagined failure. *Journal of Personality and Social Psychology, 101*, 1157. https://doi.org/10.1037/a0026105

Lieberman, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology, 58*, 259. https://doi.org/10.1146/annurev.psych.58.110405.085654

Lieberman, M. D., Straccia, M. A., Meyer, M. L., Du, M., & Tan, K. M. (2019). Social, self, (situational), and affective processes in medial prefrontal cortex (MPFC): Causal, multivariate, and reverse inference evidence. *Neuroscience and Biobehavioral Reviews, 99*, 311. https://doi.org/10.1016/j.neubiorev.2018.12.021

Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Wheelwright, S. J., Sadek, S. A., Suckling, J., … Williams, S. C. (2010). Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience, 22*, 1623. https://doi.org/10.1162/jocn.2009.21287

Macrae, C. N., Moran, J. M., Heatherton, T. F., Banfield, J. F., & Kelley, W. M. (2004). Medial prefrontal activity predicts memory for self. *Cerebral Cortex, 14*, 647. https://doi.org/10.1093/cercor/bhh025

Mar, R. A. (2010). The neural bases of social cognition and story comprehension. *SSRN, 62*, 103. https://doi.org/10.1146/annurev-psych-120709-145406

Marchlewska, M., & Cichocka, A. (2017). An autobiographical gateway: Narcissists avoid first-person visual perspective while retrieving self-threatening memories. *Journal of Experimental Social Psychology, 68*, 157. https://doi.org/10.1016/j.jesp.2016.06.003

Maresh, E. L., Allen, J. P., & Coan, J. A. (2014). Increased default mode network activity in socially anxious individuals during reward processing. *Biology of Mood & Anxiety Disorders, 4*(1), 7. https://doi.org/10.1186/2045-5380-4-7

Maresh, E. L., Teachman, B. A., & Coan, J. A. (2017). Are you watching me? Interacting effects of fear of negative evaluation and social context on cognitive performance. *Journal of Experimental Psychopathology, 8*, 303. https://doi.org/10.5127/jep.059516

Mars, R. B., Neubert, F.-X., Noonan, M. P., Sallet, J., Toni, I., & Rushworth, M. F. S. (2012). On the relationship between the "default mode network" and the "social brain". *Frontiers in Human Neuroscience, 6*, 189. https://doi.org/10.3389/fnhum.2012.00189

McIsaac, H. K., & Eich, E. (2002). Vantage point in episodic memory. *Psychonomic Bulletin and Review, 9*, 146. https://doi.org/10.3758/BF03196271

McLaughlin, K. A., & Nolen-Hoeksema, S. (2011). Rumination as a transdiagnostic factor in depression and anxiety. *Behaviour Research and Therapy, 49*, 186. https://doi.org/10.1016/j.brat.2010.12.006

Mead, G. H. (1934). *Mind, self, and society: From the standpoint of a social behaviorist* (Vol. 44, p. 587). Chicago, IL: University of Chicago Press. https://doi.org/10.2307/2179928

Michael, T., Halligan, S. L., Clark, D. M., & Ehlers, A. (2007). Rumination in posttraumatic stress disorder. *Depression and Anxiety, 24*, 307. https://doi.org/10.1002/da.20228

Mitchell, J. P. (2006). Mentalizing and Marr: An information processing approach to the study of social cognition. *Brain Research, 1079*, 66. https://doi.org/10.1016/j.brainres.2005.12.113

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50*, 655. https://doi.org/10.1016/j.neuron.2006.03.040

Moberly, N. J., & Watkins, E. R. (2008). Ruminative self-focus and negative affect: An experience sampling study. *Journal of Abnormal Psychology, 117*, 314. https://doi.org/10.1037/0021-843X.117.2.314

Moran, J. M., Heatherton, T. F., & Kelley, W. M. (2009). Modulation of cortical midline structures by implicit and explicit self-relevance evaluation. *Social Neuroscience, 4*(3), 197–211. https://doi.org/10.1080/17470910802250519

Morrison, A. S., Mateen, M. A., Brozovich, F. A., Zaki, J., Goldin, P. R., Heimberg, R. G., & Gross, J. J. (2016). Empathy for positive and negative emotions in social anxiety disorder. *Behaviour Research and Therapy, 87*, 232. https://doi.org/10.1016/j.brat.2016.10.005

Moscovitch, D. A., Orr, E., Rowa, K., Reimer, S. G., & Antony, M. M. (2009). In the absence of rose-colored glasses: Ratings of self-attributes and their differential certainty and importance across multiple dimensions in social phobia. *Behaviour Research and Therapy, 47*, 66. https://doi.org/10.1016/j.brat.2008.10.007

Moulton, S. T., & Kosslyn, S. M. (2009). Imagining predictions: Mental imagery as mental emulation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*, 1273. https://doi.org/10.1098/rstb.2008.0314

Müller-Pinzler, L., Gazzola, V., Keysers, C., Sommer, J., Jansen, A., Frässle, S., … Krach, S. (2015). Neural pathways of embarrassment and their modulation by social anxiety. *NeuroImage, 119*, 252. https://doi.org/10.1016/j.neuroimage.2015.06.036

Murray, R. J., Schaer, M., & Debbané, M. (2012). Degrees of separation: A quantitative neuroimaging meta-analysis investigating self-specificity and shared neural activation between self- and other-reflection. *Neuroscience and Biobehavioral Reviews, 36*, 1043. https://doi.org/10.1016/j.neubiorev.2011.12.013

Nelis, S., Debeer, E., Holmes, E. A., & Raes, F. (2013). Dysphoric students show higher use of the observer perspective in their retrieval of positive versus negative autobiographical memories. *Memory, 21*, 423. https://doi.org/10.1080/09658211.2012.730530

Ng, A. S., Abbott, M. J., & Hunt, C. (2014). The effect of self-imagery on symptoms and processes in social anxiety: A systematic review. *Clinical Psychology Review, 34*, 620. https://doi.org/10.1016/j.cpr.2014.09.003

Nicolle, A., Klein-Flügge, M. C., Hunt, L. T., Vlaev, I., Dolan, R. J., & Behrens, T. E. J. (2012). An agent independent axis for executed and modeled choice in medial prefrontal cortex. *Neuron, 75*, 1114. https://doi.org/10.1016/j.neuron.2012.07.023

Nigro, G., & Neisser, U. (1983). Point of view in personal memories. *Cognitive Psychology, 15*, 467. https://doi.org/10.1016/0010-0285(83)90016-6

Nordgren, L. F., Banas, K., & MacDonald, G. (2011). Empathy gaps for social pain: Why people underestimate the pain of social suffering. *Journal of Personality and Social Psychology, 100*, 120. https://doi.org/10.1037/a0020938

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain-A meta-analysis of imaging studies on the self. *NeuroImage, 31*, 440. https://doi.org/10.1016/j.neuroimage.2005.12.002

Ochsner, K. N., Ray, R. D., Cooper, J. C., Robertson, E. R., Chopra, S., Gabrieli, J. D. E., & Gross, J. J. (2004). For better or for worse: Neural systems supporting the cognitive down- and up-regulation of negative emotion. *NeuroImage, 23*, 483. https://doi.org/10.1016/j.neuroimage.2004.06.030

Ochsner, K. N., Zaki, J., Hanelin, J., Ludlow, D. H., Knierim, K., Ramachandran, T., … Mackey, S. C. (2008). Your pain or mine? Common and distinct neural systems supporting the perception of pain in self and other. *Social Cognitive and Affective Neuroscience, 3*, 144. https://doi.org/10.1093/scan/nsn006

Oosterwijk, S., Snoek, L., Rotteveel, M., Barrett, L. F., & Steven Scholte, H. (2017). Shared states: Using MVPA to test neural overlap between self-focused emotion imagery and other-focused emotion understanding. *Social Cognitive and Affective Neuroscience, 12*, 1025. https://doi.org/10.1093/scan/nsx037

Osman, S., Cooper, M. J., Hackman, A., & Veale, D. (2004). Spontaneously occurring images and early memories in people with body dysmorphic disorder. *Memory, 12*, 428. https://doi.org/10.1080/09658210444000043

Peterburs, J., Sandrock, C., Miltner, W. H. R., & Straube, T. (2016). Look who’s judging-feedback source modulates brain activation to performance feedback in social anxiety. *NeuroImage, 133*, 430. https://doi.org/10.1016/j.neuroimage.2016.03.036

Peters, U. (2016). Human thinking, shared intentionality, and egocentric biases. *Biology and Philosophy, 31*, 299. https://doi.org/10.1007/s10539-015-9512-0

Plana, I., Lavoie, M. A., Battaglia, M., & Achim, A. M. (2014). A meta-analysis and scoping review of social cognition performance in social phobia, posttraumatic stress disorder and other anxiety disorders. *Journal of Anxiety Disorders, 28*, 169. https://doi.org/10.1016/j.janxdis.2013.09.005

Plasencia, M. L., Alden, L. E., & Taylor, C. T. (2011). Differential effects of safety behaviour subtypes in social anxiety disorder. *Behaviour Research and Therapy, 49*, 665. https://doi.org/10.1016/j.brat.2011.07.005

Potheegadoo, J., Berna, F., Cuervo-Lombard, C., & Danion, J. M. (2013). Field visual perspective during autobiographical memory recall is less frequent among patients with schizophrenia. *Schizophrenia Research, 150*, 88. https://doi.org/10.1016/j.schres.2013.07.035

Pronin, E., & Ross, L. (2006). Temporal differences in trait self-ascription: When the self is seen as an other. *Journal of Personality and Social Psychology, 90*, 197. https://doi.org/10.1037/0022-3514.90.2.197

Qin, P., & Northoff, G. (2011). How is our self related to midline regions and the default-mode network? *NeuroImage, 57*, 1221. https://doi.org/10.1016/j.neuroimage.2011.05.028

Raffaelli, Q., Wilcox, R., & Andrews-Hanna, J. R. (2020). The neuroscience of imaginative thought: An integrative framework. In A. Abraham (Ed.), *The Cambridge handbook of the imagination* (pp. 332-353). Cambridge University Press.

Rothmayr, C., Sodian, B., Hajak, G., Döhnel, K., Meinhardt, J., & Sommer, M. (2011). Common and distinct neural networks for false-belief reasoning and inhibitory control. *NeuroImage, 56*, 1705. https://doi.org/10.1016/j.neuroimage.2010.12.052

Roy, M., Shohamy, D., & Wager, T. D. (2012). Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in Cognitive Sciences., 16*, 147. https://doi.org/10.1016/j.tics.2012.01.005

Ruby, P., & Decety, J. (2001). Effect of subjective perspective taking during simulation of action: A PET investigation of agency. *Nature Neuroscience, 4*, 546. https://doi.org/10.1038/87510

Samson, D., Apperly, I. A., Kathirgamanathan, U., & Humphreys, G. W. (2005). Seeing it my way: A case of a selective deficit in inhibiting self-perspective. *Brain, 128*, 1102. https://doi.org/10.1093/brain/awh464

Santiesteban, I., Banissy, M. J., Catmur, C., & Bird, G. (2012). Enhancing social ability by stimulating right temporoparietal junction. *Current Biology, 22*, 2274. https://doi.org/10.1016/j.cub.2012.10.018

Savitsky, K., Keysar, B., Epley, N., Carter, T., & Swanson, A. (2011). The closeness-communication bias: Increased egocentrism among friends versus strangers. *Journal of Experimental Social Psychology, 47*, 269. https://doi.org/10.1016/j.jesp.2010.09.005

Saxe, R. (2015). The happiness of the fish: Evidence for a common theory of one's own and others' actions. In K. D. Markman, W. M. P. Klein, & J. A. Suhr (Eds.), *Handbook of imagination and mental simulation*. Brighton, UK: Psychology Press. https://doi.org/10.4324/9780203809846.ch17

Schacter, D. L., Benoit, R. G., De Brigard, F., & Szpunar, K. K. (2015). Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions. *Neurobiology of Learning and Memory, 117*, 14. https://doi.org/10.1016/j.nlm.2013.12.008

Schilbach, L., Eickhoff, S. B., Rotarska-Jagiela, A., Fink, G. R., & Vogeley, K. (2008). Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the "default system" of the brain. *Consciousness and Cognition, 17*(2), 457–467.

Sebastian, C. L., Fontaine, N. M. G., Bird, G., Blakemore, S. J., De Brito, S. A., Mccrory, E. J. P., & Viding, E. (2012). Neural processing associated with cognitive and affective theory of mind in adolescents and adults. *Social Cognitive and Affective Neuroscience, 7*, 53. https://doi.org/10.1093/scan/nsr023

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., … Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience, 27*(9), 2349–2356. Retrieved from http://www.jneurosci.org/content/27/9/2349.short

Silani, G., Lamm, C., Ruff, C. C., & Singer, T. (2013). Right supramarginal gyrus is crucial to overcome emotional egocentricity bias in social judgments. *Journal of Neuroscience, 33*, 15466. https://doi.org/10.1523/JNEUROSCI.1488-13.2013

Spreng, R. N., & Andrews-Hanna, J. R. (2015). The default network and social cognition. In A. W. Toga (Ed.), *Brain mapping: An encyclopedic reference* (Vol. 1316, pp. 165–169). Amsterdam, Netherlands: Academic Press. https://doi.org/10.1111/nyas.12360

Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience, 21*, 489. https://doi.org/10.1162/jocn.2008.21029

Spurr, J. M., & Stopa, L. (2003). The observer perspective: Effects on social anxiety and performance. *Behaviour Research and Therapy, 41*, 1009. https://doi.org/10.1016/S0005-7967(02)00177-8

St. Jacques, P. L., Carpenter, A. C., Szpunar, K. K., & Schacter, D. L. (2018). Remembering and imagining alternative versions of the personal past. *Neuropsychologia, 110*, 170. https://doi.org/10.1016/j.neuropsychologia.2017.06.015

St. Jacques, P. L., Szpunar, K. K., & Schacter, D. L. (2017). Shifting visual perspective during retrieval shapes autobiographical memories. *NeuroImage, 148*, 103. https://doi.org/10.1016/j.neuroimage.2016.12.028

Steinbeis, N. (2016). The role of self-other distinction in understanding others' mental and emotional states: Neurocognitive mechanisms in children and adults. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 371*(1686), 20150074. https://doi.org/10.1098/rstb.2015.0074

Steinbeis, N., Bernhardt, B. C., & Singer, T. (2015). Age-related differences in function and structure of rSMG and reduced functional connectivity with DLPFC explains heightened emotional egocentricity bias in childhood. *Social Cognitive and Affective Neuroscience, 10*, 302. https://doi.org/10.1093/scan/nsu057

Steinbeis, N., & Singer, T. (2014). Projecting my envy onto you: Neurocognitive mechanisms of an offline emotional egocentricity bias. *NeuroImage, 102*, 370–380. https://doi.org/10.1016/j.neuroimage.2014.08.007

Stern, C., & West, T. V. (2016). Ideological differences in anchoring and adjustment during social inferences. *Personality and Social Psychology Bulletin, 42*, 1466. https://doi.org/10.1177/0146167216664058

Stopa, L., & Clark, D. M. (1993). Cognitive processes in social phobia. *Behaviour Research and Therapy, 31*, 255. https://doi.org/10.1016/0005-7967(93)90024-O

Suddendorf, T., Addis, D. R., & Corballis, M. C. (2009). Mental time travel and the shaping of the human mind. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*, 1317. https://doi.org/10.1098/rstb.2008.0301

Sutin, A. R., & Robins, R. W. (2008). When the “I” looks at the “Me”: Autobiographical memory, visual perspective, and the self. *Consciousness and Cognition, 17*, 1386. https://doi.org/10.1016/j.concog.2008.09.001

Sutterby, S. R., Bedwell, J. S., Passler, J. S., Deptula, A. E., & Mesa, F. (2012). Social anxiety and social cognition: The influence of sex. *Psychiatry Research, 197*, 242. https://doi.org/10.1016/j.psychres.2012.02.014

Svoboda, E., McKinnon, M. C., & Levine, B. (2006). The functional neuroanatomy of autobiographical memory: A meta-analysis. *Neuropsychologia, 44*, 2189. https://doi.org/10.1016/j.neuropsychologia.2006.05.023

Tamir, D. I., Bricker, A. B., Dodell-Feder, D., & Mitchell, J. P. (2015). Reading fiction and reading minds: The role of simulation in the default network. *Social Cognitive and Affective Neuroscience, 11*, 215. https://doi.org/10.1093/scan/nsv114

Tamir, D. I., & Mitchell, J. P. (2010). Neural correlates of anchoring-and-adjustment during mentalizing. *Proceedings of the National Academy of Sciences, 107*, 10827. https://doi.org/10.1073/pnas.1003242107

Tamir, D. I., & Mitchell, J. P. (2013). Anchoring and adjustment during social inferences. *Journal of Experimental Psychology: General, 142*, 151. https://doi.org/10.1037/a0028232

Todd, A. R., Forstmann, M., Burgmer, P., Brooks, A. W., & Galinsky, A. D. (2015). Anxious and egocentric: How specific emotions influence perspective taking. *Journal of Experimental Psychology: General, 144*, 374. https://doi.org/10.1037/xge0000048

Treynor, W., Gonzalez, R., & Nolen-Hoeksema, S. (2003). Rumination reconsidered: A psychometric analysis. *Cognitive Therapy and Research, 27*, 247. https://doi.org/10.1023/A:1023910315561

Valenti, G., Libby, L. K., & Eibach, R. P. (2011). Looking back with regret: Visual perspective in memory images differentially affects regret for actions and inactions. *Journal of Experimental Social Psychology, 47*, 730. https://doi.org/10.1016/j.jesp.2011.02.008

van der Heiden, L., Scherpiet, S., Konicar, L., Birbaumer, N., & Veit, R. (2013). Inter-individual differences in successful perspective taking during pain perception mediates emotional responsiveness in self and others: An fMRI study. *NeuroImage, 65*, 387. https://doi.org/10.1016/j.neuroimage.2012.10.003

van der Meer, L., Costafreda, S., Aleman, A., & David, A. S. (2010). Self-reflection and the brain: A theoretical review and meta-analysis of neuroimaging studies with implications for schizophrenia. *Neuroscience and Biobehavioral Reviews, 34*, 935. https://doi.org/10.1016/j.neubiorev.2009.12.004

Van der Meer, L., Groenewold, N. A., Nolen, W. A., Pijnenborg, M., & Aleman, A. (2011). Inhibit yourself and understand the other: Neural basis of distinct processes underlying theory of mind. *NeuroImage, 56*, 2364. https://doi.org/10.1016/j.neuroimage.2011.03.053

van der Weiden, A., Prikken, M., & van Haren, N. E. M. (2015). Self-other integration and distinction in schizophrenia: A theoretical analysis and a review of the evidence. *Neuroscience and Biobehavioral Reviews, 57*, 220. https://doi.org/10.1016/j.neubiorev.2015.09.004

Varela, F. J., Thompson, E., & Rosch, E. (2017). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.

Vazire, S., & Carlson, E. N. (2010). Self-knowledge of personality: Do people know themselves? *Social and Personality Psychology Compass, 4*, 605. https://doi.org/10.1111/j.1751-9004.2010.00280.x

Wallace, H., & Tice, D. (2012). Reflected appraisal through a 21st-century looking glass. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity*. New York, NY: Guilford Press. https://doi.org/10.1109/CAR.2010.5456542

Wallin, A. (2011). Is egocentric bias evidence for simulation theory? *Synthese, 178*, 503. https://doi.org/10.1007/s11229-009-9653-2

Wang, T., Yang, L. L., Yang, Z., & Huang, X. T. (2019). Imagining my painful hand is not mine: Self-distancing relieves experimental acute pain induced by a cold pressor task. *Journal of Pain, 20*, 358. https://doi.org/10.1016/j.jpain.2018.10.001

Washburn, D., Wilson, G., Roes, M., Rnic, K., & Harkness, K. L. (2016). Theory of mind in social anxiety disorder, depression, and comorbid conditions. *Journal of Anxiety Disorders, 37*, 71. https://doi.org/10.1016/j.janxdis.2015.11.004

Watkins, E., & Teasdale, J. D. (2004). Adaptive and maladaptive self-focus in depression. *Journal of Affective Disorders, 82*, 1. https://doi.org/10.1016/j.jad.2003.10.006

Wells, A., Clark, D. M., & Ahmad, S. (1998). How do I look with my minds eye: Perspective taking in social phobic imagery. *Behaviour Research and Therapy, 36*, 631. https://doi.org/10.1016/S0005-7967(98)00037-0

Wild, J., Hackmann, A., & Clark, D. M. (2008). Rescripting early memories linked to negative images in social phobia: A pilot study. *Behavior Therapy, 39*, 47. https://doi.org/10.1016/j.beth.2007.04.003

Willard, D. F. X., & Markman, A. B. (2017). Anchoring on self and others during social inferences. *Topics in Cognitive Science, 9*, 819. https://doi.org/10.1111/tops.12275

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology, 55*, 493. https://doi.org/10.1146/annurev.psych.55.090902.141954

Yankouskaya, A., Humphreys, G., Stolte, M., Stokes, M., Moradi, Z., & Sui, J. (2017). An anterior-posterior axis within the ventromedial prefrontal cortex separates self and reward. *Social Cognitive and Affective Neuroscience, 12*, 1859. https://doi.org/10.1093/scan/nsx112

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., … Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology, 106*(3), 1125–1165. https://doi.org/10.1152/jn.00338.2011

Zaki, J., & Ochsner, K. (2009). The need for a cognitive neuroscience of naturalistic social cognition. *Annals of the New York Academy of Sciences, 1167*, 16. https://doi.org/10.1111/j.1749-6632.2009.04601.x

The Self–Other Distinction in Psychopathology: Recent Developments from a Mentalizing Perspective

Patrick Luyten, Celine De Meulemeester, and Peter Fonagy

Introduction

Over the past decades, the mentalizing approach to the development and treatment of psychopathology has increased in popularity, as is testified by the wide-ranging research included in this book. Mentalizing, or reflective functioning, refers to the human capacity to understand oneself and others in terms of internal mental states. It is a species-specific capacity that is present in higher primates, and in humans specifically, and is largely or completely absent in other species (Tomasello, 2018; Tomasello & Vaish, 2013). It plays a key role in the capacity of humans to navigate their complex interpersonal world. Given its centrality in human functioning, it is unsurprising that studies have amply demonstrated that most if not all forms of psychopathology are characterized by temporary or chronic impairments in mentalizing (Fonagy & Luyten, 2016). There is also increasing evidence for the effectiveness and cost-effectiveness of psychosocial treatments that focus on improving mentalizing (Blankers et al., 2019; Smits et al., 2019; Volkert, Hauschild, & Taubner,

P. Luyten (✉)
Faculty of Psychology and Educational Sciences, University of Leuven, Leuven, Belgium

Research Department of Clinical, Educational and Health Psychology, University College London, London, UK
e-mail: patrick.luyten@kuleuven.be

C. De Meulemeester
Faculty of Psychology and Educational Sciences, University of Leuven, Leuven, Belgium
e-mail: celine.demeulemeester@kuleuven.be

P. Fonagy
Research Department of Clinical, Educational and Health Psychology, University College London, London, UK

Psychology and Language Sciences, University College London, London, UK
e-mail: p.fonagy@ucl.ac.uk

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_34

2019). Mentalizing may therefore be an important transtheoretical and transdiagnostic concept in explaining vulnerability to psychopathology and its treatment.

Both developmental psychopathology and neuroscience studies have shown that mentalizing is a multidimensional capacity that is underpinned by four dimensions: (a) automatic–controlled, (b) internal–external, (c) self–other, and (d) cognitive–affective (Lieberman, 2007; Luyten & Fonagy, 2015). Different types of mental disorder appear to be associated with different types of imbalances between these various dimensions, resulting in *mentalizing profiles* that are characteristic of each disorder. For instance, there is now good evidence (Fonagy & Luyten, 2016) to suggest that borderline personality disorder (BPD) is characterized by a rapid loss of controlled mentalizing in high-arousal contexts, leading to overreliance on fast, automatic, and biased mentalizing that is dominated by affective experiences at the expense of cognitive control. Moreover, this tendency for fast and biased mentalizing seems to be strongly associated with the rapid (over)interpretation of external social cues (such as facial expressions of others) at the expense of a focus on mental interiors of others and, as a result, a tendency to conflate the mental states of self and others (also known as identify diffusion). This tendency to conflate self and other would also explain the increased susceptibility for emotional contagion of individuals with BPD, as the emotions of others may be easily misunderstood as originating from the self (Fonagy & Luyten, 2016).

In this chapter, we discuss emerging knowledge concerning the neural circuits underlying these dimensions in mentalizing and their role in explaining psychopathology, with a focus on problems with the self–other dimension. Indeed, problems with distinguishing between self and others are implicated in a wide variety of mental disorders, such as depression, personality disorders, and psychosis (Beck, Freeman, & Davis, 2004; Fonagy, Luyten, & Allison, 2015; Kernberg & Caligor, 2005). In DSM-5 (American Psychiatric Association, 2013), impairments in the self are considered to be one of the central features of personality disorders. The chapter focuses on the following question: How can the mentalizing approach to psychopathology contribute to our understanding of these problems with the self–other distinction (SOD)? First, we review the four dimensions underlying mentalizing, the neural circuits involved in them, and their role in different types of psychopathology. Next, we focus on the phenomenology of self–other confusion. In the final section, we review recent neuroscience studies investigating these issues in BPD and other mental disorders.

Dimensions of Mentalizing

Neuroscientific and behavioral studies together suggest that mentalizing can be organized around four dimensions, each of which has relatively distinct underlying neural circuits. First and foremost, mentalizing can be relatively fast, reflexive and *automatic*, or more *controlled*, explicit and deliberate (Satpute & Lieberman, 2006). Automatic or implicit mentalizing is subserved by phylogenetically older brain networks, while controlled or explicit mentalizing is underpinned by evolutionary newer brain circuits that are shaped throughout development (Lieberman, 2007; Satpute & Lieberman, 2006). In situations of increasing physiological arousal or stress, a switch occurs from controlled to automatic mentalizing (Lieberman, 2007; Mayes, 2006). Although this switch clearly has survival advantages in acutely threatening circumstances, overreliance on automatic mentalizing may result in biased assumptions about the self and others in more complex social interactions. More controlled reflection is often required to meet the high demands of the social world, in terms of communication, collaboration, and competition, making social interactions especially challenging for those individuals who have a low "switch point" for the change from automatic to controlled mentalizing under stress (Fonagy et al., 2015). Both biological (i.e., capacity for effortful control) and environmental (i.e., attachment history) factors, and their interactions, are known to contribute to inter-individual differences in the capacity for controlled mentalizing (for a comprehensive review, see Long, Verbeke, Ein-Dor, & Vrticka, 2019).

Furthermore, mentalizing can be either *internally based*, that is, focused on the mental interiors of the self and others, or *externally based*, focusing on cues such as facial expressions or tone of voice to infer mental states. Internally based mentalizing recruits the medial frontoparietal network, and is thus more controlled and reflective than externally focused mentalizing, which relies more on the lateral frontotemporoparietal network (Lieberman, 2007). An excessive focus on external features to infer mental states thus harbors the risk of jumping to conclusions.

Reflecting on self and others requires *cognitive* skills such as perspective-taking (i.e., the capacity to inhibit one's own mental representations in order to take the perspective of another person) and belief-desire reasoning, but it is equally important for mentalizing to be grounded in an embodied *affective* reality. Whereas cognitive mentalizing again relies on controlled processing, affective mentalizing may, at least at the basic neural level, be largely automatic and embodied (Sabbagh, 2004). Researchers investigating empathy have identified integrated but clearly dissociable cognitive and affective routes to knowing others (Shamay-Tsoory & Aharon-Peretz, 2007; Stietz, Jauk, Krach, & Kanske, 2019; Uribe et al., 2019). Indeed, the capacity for empathy entails a more basic "emotional contagion"—that is, feeling another person's emotions as if they are one's own—as well as the more reflective capacity for perspective-taking, which crucially entails awareness that another person is the source of one's emotions (i.e., requiring SOD) (Kanske, 2018).

This brings us to the discussion of the neural circuitry underpinning the capacity to mentalize the *self* and *others*. Self–other processing is underpinned by overlapping neural networks (Decety & Sommerville, 2003). Indeed, brain activation did not differ when reflecting on the self or the other in either BPD patients or controls (Beeney, Hallquist, Ellison, & Levy, 2016). Patients with schizophrenia recruited even more overlapping brain maps for self and non-self, especially in the inferior parietal lobule (IPL), which in controls is active only when representing the non-self (Jardri et al., 2011). Two large networks have been identified as being implicated in self–other processing (Ripoll, Snyder, Steele, & Siever, 2013). The *shared representation* (SR) system is a bodily based, frontoparietal mirror neuron network allowing for the sharing of others' mental states through sensorimotor simulation. This system allows the individual to know how others feel "from the inside," as neural activation is similar while experiencing states of mind and observing others experiencing the same states of mind. The SR system recruits the inferior frontal gyrus (IFG), IPL (both of these zones are rich in mirror neurons), anterior insula, and anterior cingulate cortex (ACC) (both of which are involved in observed and felt pain) (Lombardo, Chakrabarti, & Baron-Cohen, 2009). The second system, *mental state attribution* (MSA), allows the individual to reflect upon the mental states of both self and others in more abstract and symbolic ways, and is mainly shaped by interpersonal relationships. It has also been found in primates, and in humans it fully develops only in adolescence (Lackner, Bowman, & Sabbagh, 2010). It involves a cortical midline system consisting of the ventromedial prefrontal cortex (vmPFC), dorsomedial prefrontal cortex (dmPFC), temporoparietal junction (TPJ), medial temporal pole, and perhaps also the posterior cingulate cortex (Lieberman, 2007; Uddin, Iacoboni, Lange, & Keenan, 2007).

The overlap in brain networks for self–other processing, as well as the remarkable human capacity for sharing emotions through SR processing, raises the question of how self and other can be distinguished (Bird & Viding, 2014). Indeed, the sharing of mental states through embodied simulation has incredible advantages for empathy, collaboration, and mentalizing, but also holds the potential for conflating the mental states of self and others. Self–other conflation may arise when one misinterprets the embodied simulation of others' mental states as originating from the self (altercentric bias), or when one assumes to understand the mind of the other based on one's own experience (egocentric bias) (Silani, Lamm, Ruff, & Singer, 2013). The realization that the human neuroarchitecture inevitably entails the possibility of misunderstanding others has become a key guiding principle of mentalization-based treatments. It is clear that SOD—the capacity to distinguish between the mental states of self and other—is necessary to overcome this risk and to create both self-awareness and awareness of others (Tsakiris, 2017).

The neural mechanism of SOD is not yet fully understood. Areas of the MSA system such as the TPJ and the medial prefrontal cortex (mPFC) have been implicated in the inhibition of automatic imitation (Brass, Ruby, & Spengler, 2009; Sowden & Catmur, 2015), suggesting that controlled reflection upon mental states (i.e., MSA) allows inhibition of the automatic imitation of others' mental states (i.e., SR) and fosters SOD. However, a recent meta-analysis suggests that general cognitive control networks (e.g., right IPL, the right supramarginal gyrus (rSMG), and the right superior temporal gyrus (rSTG)) are implicated in inhibition of imitation rather than the MSA system, challenging this idea (Darda & Ramsey, 2019). Nonetheless, the TPJ has been implicated in many measures of SOD (Bardi, Six, & Brass, 2017; Eddy, 2016; Heinisch, Kruger, & Brune, 2012), with implications for social functioning. For instance, stimulation of the TPJ with transcranial direct current stimulation (tDCS) improved SOD in imitation and perspective-taking tasks (Santiesteban, Banissy, Catmur, & Bird, 2012), and the inhibition of the right TPJ using repetitive transcranial magnetic resonance (rTMS) increased the tendency to attribute hostile intent (Giardina, Caltagirone, & Oliveri, 2011). The TPJ has been related to reorienting attention in general (Igelstrom, Webb, Kelly, & Graziano, 2016; Schurz, Tholen, Perner, Mars, & Sallet, 2017) and toward a social partner in particular (Gvirts & Perlmutter, 2019; Krall et al., 2016), and so its role in SOD may be to reorient focus from the self to the other to "tune into" the other. Furthermore, the inferior frontal gyrus (IFG) has likewise been implicated in SOD (Backasch et al., 2014) and is hypothesized to be involved in the comparison of interoceptive (i.e., arising from within the body) and exteroceptive (i.e., arising from outside the body) signals to determine whether an action was caused by the self or the environment. Although the identification of the neural mechanism for SOD is preliminary, research has routinely implicated the TPJ and the PFC, suggesting that distinguishing self from others is an effortful process requiring inhibitory control, the reorientation of attention, and the comparison of match–mismatch between interoceptive and exteroceptive cues.

The Self–Other Distinction in Psychopathology

The self and its impairments are multidimensional, with a distinction commonly being made between the narrative or extended self, that is, the higher-order mental and symbolic representation of the self, and the core or minimal self, which includes the lower-level senses of body ownership and of agency (Gallagher, 2000; Zahavi, 2010). The sense of ownership over a coherent body that is distinct from the environment and the sense that one is the cause of one's actions (agency) are crucial for distinguishing self from other (Braun et al., 2018; Fotopoulou & Tsakiris, 2017; Kahl & Kopp, 2018).

Impairments in the sense of agency or self-directedness, also referred to as identity diffusion, have consistently been identified as a key feature of psychopathology (Adler, Chin, Kolisetty, & Oltmanns, 2012; Barnow, Ruge, Spitzer, & Freyberger, 2005; Bender & Skodol, 2007; Jørgensen et al., 2012; Richetin, Preti, Costantini, & De Panfilis, 2017). Self-impairments are notably pronounced in personality disorders, particularly BPD (Fuchs, 2007; Sollberger et al., 2012; Wilkinson-Ryan & Westen, 2000) and have been recognized as a central diagnostic dimension of personality disorders in Sect. III of the DSM-5 (Bender, Morey, & Skodol, 2011). Impairments in the sense of self and identity are also prevalent in depression and anxiety, in the form of feelings of derealization and depersonalization, but also in identifying with criticism from others (Luyten & Fonagy, 2018). Of course, in psychosis, severe mentalizing impairments can lead to serious distortions in the experience of self and/or others (Debbané et al., 2016). This self–other confusion can be so profound that self-generated representations may be experienced as originating from the environment. For instance, when hearing voices, the person actually believes they are coming from another person or entity, or they have the feeling that their own body movements are controlled by an external agent (van der Weiden, Prikken, & van Haren, 2015).

The emergence of the self is biologically predisposed but to an important extent impacted by interpersonal relationships, especially those proximal, bodily interactions with caregivers, in which they carefully mirror the infant's body and mental states (Fonagy, Gergely, Jurist, & Target, 2002; Fonagy, Gergely, & Target, 2007). It is therefore not surprising that self-impairments have been closely associated with problems with the capacity to form and maintain relationships (Beeney et al., 2019; De Meulemeester, Lowyck, Vermote, Verhaest, & Luyten, 2017; Lowyck, Luyten, Verhaest, Vandeneede, & Vermote, 2013), with high levels of impulsivity, feelings of dissociation, self-injury, and a strong sense of inner pain prompted by experiences of rejection, isolation, or abandonment (Gunderson & Lyons-Ruth, 2008; Yen et al., 2004). Indeed, when confronted with feelings of emptiness, that is, a lack of continuity in the experience of self (and as a result others), often intense feelings of despair, inner pain, helplessness, and hopelessness emerge. In an attempt to achieve a sense of control and relieve of such unbearable feelings, one may resort to self-harm ("I will show others how bad and empty I am") and/or engage in impulsive behaviors that harm the self or others (e.g., promiscuity in an attempt to counter feelings of emptiness, or the use of drugs to dampen these feelings).

From a mentalizing perspective, the experience of self-coherence and self-continuity is an illusion (Bargh, 2011, 2014) that is the product of the capacity for mentalizing (Han & Northoff, 2009; Northoff et al., 2006). Stated otherwise: the self (and feelings of self-coherence and self-continuity) is always created "on-line" and in contrast with what is "not me." In every moment, mental states of self and those of others need to be co-represented and at the same time distinguished, which may be particularly difficult in high-arousal contexts such as social interactions (Deckers et al., 2015; Nolte et al., 2013). BPD patients seem to have severe problems in this process, resulting in a paradoxical combination of self–other diffusion, instability, and fluidity, but at the same time marked rigidity (Fonagy et al., 2015). Indeed, the syndrome of identity diffusion inherently entails being very impressionable and sensitive to signals from the social environment, and fearing the risk of "losing one's self" in relationships with others (Bender & Skodol, 2007; Jørgensen, 2006). It is as if the fragile self-representation of individuals with BPD is easily "overridden" by the representation of the other, resulting in marked instability in their sense of self in response to ever-changing environmental (i.e., exteroceptive) input.

However, individuals with BPD are at the same time characterized by marked rigidity, or a lack of flexibility to adopt positions other than the perspective they hold at a given moment. So, despite their interpersonal sensitivity, individuals with BPD are notoriously "hard to reach" and have a hampered capacity for change and social learning, as expressed in their high levels of epistemic mistrust (Fonagy et al., 2015; Luyten & Fonagy, 2018). Social learning crucially requires SOD because one needs to be able to represent the other as having a separate mind that can differ from one's own. Reduced perspective-taking, in the sense of reduced ability to adopt another person's perspective, is typical of BPD patients (Haas & Miller, 2015; Harari, Shamay-Tsoory, Ravid, & Levkovitz, 2010; New et al., 2012), and they have also been found to have more difficulty in mentalizing from an altercentric, third-person perspective (Colle et al., 2019). Additionally, for genuine change to occur, newly acquired social information needs to be incorporated into the self; in other words, it needs to be evaluated against and integrated into existing self-knowledge and autobiographical memory (Fonagy et al., 2015). In our view, learning from others crucially requires flexibility in representing the "other" ("What does this other person want to teach me?") and "self" ("Is this information relevant for me? How do I incorporate this information with what I already know?"). In this way, rather than being processed episodically (e.g., "My therapist told me about X last week"), social information can be processed in terms of its generalizability and usefulness to the self (e.g., "Our discussion last week made me change my views on X (or not)"). A recent functional magnetic resonance imaging (fMRI) study suggested that BPD patients have difficulty doing this (van Schie, Chiu, Rombouts, Heiser, & Elzinga, 2019). In our view, self–other rigidity is detrimental for change and learning, as a rigid focus on "self" may hamper the consideration of the knowledge of others, while a rigid focus on "other" may leave one without the capacity to "filter out" information based on usefulness and applicability to the self and to incorporate social information into self-knowledge.

We argue that in individuals with BPD, this paradox of instability and rigidity is created by their lack of flexibility to switch between representations of self and other and to simultaneously keep both "self" and "other" in mind. Indeed, in a mind-wandering task, when participants were asked to rate their self-generated thoughts, BPD patients rated these as more extremely self-related and as extremely other-related, while controls' self-generated thoughts were more self–other ambiguous, pertaining both to self and others at the same time (Kanske et al., 2016). Tolerating self–other ambiguity and maintaining SOD in close bodily interactions may be especially challenging for these individuals (de Bézenac, Swindells, & Corcoran, 2018). Sowden and Shah (2014) argue that the control over self–other representations (i.e., focusing either on self or on other, depending on situational demands, when both are represented simultaneously) and flexibility in switching between the representations of self and others are key in navigating the demands of complex social interactions, as some interpersonal situations call for the inhibition of the self and the enhancement of the other, as in perspective-taking, while other situations call for enhancement of the self-perspective.

This idea of a lack of self–other flexibility is not new, as attachment approaches have described rigidity as a central feature of personality disorders (McWilliams, 2011). Specifically, Blatt and colleagues (Blatt & Luyten, 2009; Luyten & Blatt, 2013) have emphasized that while adaptive personality development is characterized by the capacity to constantly re-evaluate the sense of self and relatedness in the course of development, psychopathology involves a lack of the capacity to move flexibly between relatedness and self-definition, leading to an exaggerated emphasis on either identity and autonomy ("self") or attachment and relatedness ("other"). In two studies on the relationship between attachment and the content of self–other representations, anxious attachment related to overestimating self–other similarity, while attachment avoidance related to underestimated self–other similarity (Mikulincer & Horesh, 1999; Mikulincer, Orbach, & Iavnieli, 1998). This was in contrast with securely attached individuals who made more accurate estimations of self–other similarity. Furthermore, when observing their own interactions with their mothers on tape, anxiously attached adolescents rated their mothers' affect as more similar to their own, showing a decreased capacity for SOD (Diamond, Fagundes, & Butterworth, 2012).

The role of oxytocin (OT), a key neuromodulator implicated in the attachment system, in SOD is not yet clear, but emerging evidence suggests that it may foster self–other flexibility. Studies have found the administration of OT to increase tolerance for self–other ambiguity implicitly, in terms of increased non-conscious mimicking of others' mannerisms and in increased gaze behavior toward information about others when making judgements about the self (Pfundmair, Rimpel, Duffy, & Zwarg, 2018). Furthermore, the administration of OT enhanced speed and performance in experimental SOD tasks (Colonnello, Chen, Panksepp, & Heinrichs, 2013; Tomova, Heinrichs, & Lamm, 2019) and reduced self-bias in speed of labelling traits of self and others (Zhao et al., 2016). Shifting from representing "self" to representing "other" is an effortful process, and the emerging findings seem to suggest that OT may facilitate this switch, resulting in enhanced SOD.

Secure attachment relates to parent–infant interactions that are characterized by behavioral and physiological synchrony and affective attunement, which fosters self–other fusion (Biro, Alink, Huffmeijer, Bakermans-Kranenburg, & Van, 2017; Feldman, 2012). However, this synchrony is never perfect but rather "good enough" and moments of asynchrony and non-attunement inevitably arise that promote SOD (Fotopoulou & Tsakiris, 2017). Parents' mentalizing ability or "mind-mindedness" probably allows them to mirror their infant's emotional states in a way that is contingent (i.e., mirroring the *right* emotional state that corresponds with the infant's self, providing synchrony) and at the same time marked (i.e., mirroring in an exaggerated way, signalling that the parent is displaying the emotion of the infant rather than their own emotion, providing asynchrony) (Fonagy et al., 2002, 2007; Gergely & Watson, 1999; Meins, Fernyhough, Fradley, & Tuckey, 2001). This type of self–other ambiguity in parent–infant interactions in terms of providing a balance between self–other synchrony and asynchrony may provide a safe (albeit challenging, because inherently ambiguous) context for infants to learn to make SODs (de Bézenac et al., 2018). Individuals with BPD may have had limited access to such interactions, either because caregivers were physically absent and there was little opportunity for practicing SOD, and/or because caregivers did not provide a balance between synchrony and asynchrony. Instead, caregivers may have provided too much self–other asynchrony because they were not able to adequately mentalize the infant's emotional state (lack of *congruency*), and/or may have provided too much synchrony by overwhelming the infant with the own emotional states in response to the infant's emotion (lack of *marking*). This is consistent with findings of high levels of attachment insecurity (Agrawal, Gunderson, Holmes, & Lyons-Ruth, 2004) and complex trauma (Ball & Links, 2009; Chanen & Kaess, 2012; de Aquino Ferreira, Queiroz Pereira, Neri Benevides, & Aguiar Melo, 2018; Stepp, Lazarus, & Byrd, 2016) in BPD patients.

The effects of trauma on the development of feelings of self and identity, the capacity to form interpersonal relationships, and emotion regulation have been extensively demonstrated (Horowitz, 2015; Luyten, Campbell, & Fonagy, 2019; Villalta, Smith, Hickin, & Stringaris, 2018). From a mentalizing perspective, an experience becomes traumatic in the absence of "relational referencing," that is, the meaningful framing and reframing of the traumatic event with the help of another person who tries to represent your mental states and reflects them back to you in a way that is more manageable and in a way "digested" (i.e., "marked mirroring"). This process helps to recalibrate the mind, and its absence may lead to a feeling that one's mind is alone (Allen, 2012). This means that the person cannot safely access another person's mind, which is crucial for the development of the self and SOD. It is the experience of being held in mind by someone else, that is, the experience of another person trying to represent your mental states, that we see as crucial in restoring a sense of agency and control, and ultimately a sense of selfhood, in individuals who have experienced trauma. A recent study found that after 12 months of psychotherapy, BPD patients showed increased agency when narrating their life stories, showing that agency may indeed be an important mechanism for change in psychotherapy (Lind et al., 2019).

However, a singular focus on relational trauma in early parent–child relationships would be too simplistic, as peer relationships, the broader environment, and later relational experiences have been found to mediate and moderate the impact of trauma on BPD (Belsky et al., 2012; Carlson, Egeland, & Sroufe, 2009; Carlson, Sroufe, & Egeland, 2004; Salvatore, Haydon, Simpson, & Collins, 2013; Shakoor et al., 2012). Furthermore, genetic factors and temperamental differences are also implicated in vulnerability for BPD. Indeed, heritability estimates of BPD range between 40% and 50% (Bornovalova, Hicks, Iacono, & McGue, 2009; Distel et al., 2009). For instance, high levels of impulsivity/aggression and hypersensitivity to social information have been implicated in BPD (for a review, see Bateman, O'Connell, Lorenzini, Gardner, & Fonagy, 2016) and may play an important role in developmental pathways to BPD in combination with trauma (but also in its absence). For instance, individuals with BPD have been reported to show elevated levels of emotional and physical pain in response to negative experiences (Holm & Severinsson, 2008; Sansone & Sansone, 2012) and hypersensitivity to social exclusion (Bungert et al., 2015; De Panfilis, Riva, Preti, Cabrino, & Marchesi, 2015). It is as if there is a lack of boundaries between the self and the environment in BPD, which may in part be genetically predisposed (Gunderson & Lyons-Ruth, 2008).

Recent Developments in Research on Self–Other Distinction in Psychopathology

The mentalizing approach is increasingly focusing on the embodied aspect of mentalizing (Fotopoulou & Tsakiris, 2017). Indeed, it is widely believed that the "self" first emerges as a bodily experience in interaction with other proximal bodies (Gallagher, 2000; Zahavi, 2014). Comparator models of the "minimal self" (i.e., the embodied or sensorimotor self) emphasize the need to compare stimuli originating outside the body (i.e., exteroception) and inside it (i.e., interoception) in order to determine whether an event is caused by the self or by another person or the environment (Braun et al., 2018; Gallagher, 2000; Kahl & Kopp, 2018; Tsakiris, 2017). If there is a match between internally and externally derived signals, it is most likely that these have been caused by the self, whereas in the case of a mismatch, the event is attributed to an external cause. The capacity for sensorimotor SOD therefore relies on the multisensory integration of interoceptive and exteroceptive cues and their relative precision, i.e., the validity that one assigns to these cues (Fotopoulou & Tsakiris, 2017). Several experimental paradigms have been developed to investigate individuals' interoceptive accuracy and their sensitivity to exteroceptive cues, and these paradigms have recently been applied to investigate self-impairments in psychopathology. These paradigms are also of key relevance for emerging research on embodied mentalizing, as they focus on the role of interoceptive information on mentalizing the self as distinct from others.

Individuals with impaired interoceptive awareness will more readily attenuate less precise interoceptive signals when confronted with exteroceptive input, making the "self" more modifiable and hampering the experience of the self as stable and continuous over time (Palmer & Tsakiris, 2018). Indeed, impaired interoception has been related to blurred self–other boundaries (Tajadura-Jimenez & Tsakiris, 2014). Interoceptive accuracy has also been found to be reduced in individuals with schizophrenia (Ardizzi et al., 2016) and in patients with moderate but not severe depression (Eggart, Lange, Binser, Queri, & Muller-Oerlinghausen, 2019), while superior interoception has been associated with anxiety-specific arousal symptoms (Dunn et al., 2010). Interoceptive awareness has been found to be reduced in BPD, in the form of reduced amplitude of heartbeat-evoked potentials (HEPs), an indicator of the cortical representation of afferent signals from the cardiovascular system, which is related to more pronounced emotion dysregulation and smaller anterior insula grey matter volume (Muller et al., 2015). However, other studies have found maintained levels of interoceptive accuracy in BPD, which may reflect the fact that attention and confidence in one's own perception modulates interoceptive awareness (Loffler, Foell, & Bekrater-Bodmann, 2018; Muller et al., 2015) or the fact that there is a lack of a gold standard for interoceptive tasks (Palmer, Ainley, & Tsakiris, 2019). Although more research is needed, several forms of psychopathology seem to be associated with a reduced capacity to mentalize the embodied self.

Several experimental paradigms are used to investigate individuals' susceptibility to self–other blurring by manipulating exteroceptive input to match interoceptive signals. For instance, in the "rubber hand illusion" (RHI), a rubber hand is stroked in synchrony or asynchrony with the participant's hand to create illusory ownership over the rubber hand (Botvinick & Cohen, 1998). Enhanced illusory ownership over the rubber hand has been found in patients with schizophrenia (Klaver & Dijkerman, 2016), eating disorders (Eshkevari, Rieger, Longo, Haggard, & Treasure, 2012), and BPD, where illusion strength was related to dissociative symptoms (Bekrater-Bodmann et al., 2016) and trait psychoticism (Neustadter, Fineberg, Leavitt, Carr, & Corlett, 2019). Several studies using the RHI have found that synchronicity had less of an impact in populations with mental disorders, for instance, patients with schizophrenia (Prikken et al., 2019), body dysmorphic disorder (Kaplan, Enticott, Hohwy, Castle, & Rossell, 2014), or BPD (Neustadter et al., 2019), compared with controls. In these disorders, the illusion not only occurred in the synchronous condition but was maintained during asynchronous stroking. In nonclinical controls, the mismatch that occurs between the seen (exteroceptive) and felt (interoceptive) stroking in the asynchronous condition serves as a strong cue indicating that the rubber hand does not belong to the self, but individuals with self-disturbance seem less able to detect this mismatch (Kaplan et al., 2014). A similar tendency to incorporate inconsistent exteroceptive information was observed in BPD patients during a finger-tapping task that measures imitation-inhibition. While showing similar response facilitation to that of controls in the congruent condition, the BPD patients showed enhanced interference when observing an incongruently imitating hand (Hauschild et al., 2018). Other studies have also found that control over irrelevant stimuli seems to be lacking in individuals with BPD (Domes et al., 2006). Heightened precision attributed to exteroceptive stimuli, compared with relatively imprecise interoceptive signals, may account for the more malleable sense of bodily self in BPD. Although more research is needed, these findings seem to suggest that BPD is associated, not only with problems in SOD in terms of mental states, but also in their bodily sense of self, especially in relation to symptoms of dissociation (Bekrater-Bodmann et al., 2016) and psychoticism (Neustadter et al., 2019).

On a more abstract semantic level, BPD patients were found to be more affected by, and showed more right TPJ activation in response to, negative but not positive feedback, while the controls showed the opposite pattern of responses (van Schie et al., 2019) Considering the role of the TPJ in reorienting attention to the other (Dugue, Merriam, Heeger, & Carrasco, 2018; Gvirts & Perlmutter, 2019; Igelstrom et al., 2016; Krall et al., 2016; Schurz et al., 2017), it seems that BPD patients try to represent the mind of the other when the other provides negative feedback, but not when the other provides positive feedback. Furthermore, their mood ratings in response to the feedback were less modulated by the applicability of the feedback (van Schie et al., 2019). This again demonstrates the social hypersensitivity of BPD patients, and how even information that is inconsistent with self-knowledge (i.e., not applicable to the self) is incorporated into the self by these individuals. This seemed to be corroborated by the BPD patients' reduced activation of the precuneus, a brain region that may be involved in putting self-relevant stimuli into autobiographical context (Northoff et al., 2006) and that was previously found to relate to the applicability of feedback to the self (van Schie, Chiu, Rombouts, Heiser, & Elzinga, 2018). The control participants seemed to be able to weigh the incoming social information (exteroception) against their existing self-knowledge (interoception), which protected them and provided them with a filter for deciding what information to incorporate into the self ("this is me"); this mechanism may be disrupted in BPD. The process of comparing new information against self-knowledge may be crucial for social learning (Fonagy et al., 2015).

Several fMRI studies have investigated the neural networks involved in self–other processing in BPD. During passive viewing of emotional stimuli and during very basic social cognitive tasks that do not explicitly require SOD, BPD patients recruit the somatosensory and premotor cortices (SR system) to a stronger degree than controls, suggesting that individuals with BPD resonate more strongly with others' emotions (Dziobek et al., 2011; Mier et al., 2013; Schulze, Schmahl, & Niedtfeld, 2016; Sosic-Vasic et al., 2019). This stronger resonance with others crucially harbors the risk to confuse the mental states of others as originating from the self. This heightened SR activation is coupled with hypoactivation of the superior temporal sulcus (STS) and TPJ (Haas & Miller, 2015), and the IFG (Mier et al., 2013; Sosic-Vasic et al., 2019), all of which have been implicated in SOD. This suggests that under limited task instructions, individuals with BPD resonate more strongly with others and show a reduced capacity to mentalize explicitly and make SODs compared to controls.

When having received explicit instructions to make SODs, BPD patients may be able to recruit these mentalizing areas, but may overcompensate and start to hypermentalize (Beeney et al., 2016). When instructed to answer questions about personality traits of the self and a familiar other, from a first-person (e.g., "Are you kind?") or a third-person (e.g., "Does your friend think you are kind?") perspective, individuals with BPD showed hyperactivation of mentalizing (i.e., MSA) regions (e.g., the mPFC, right TPJ, and precuneus) and hypoactivation of sensory, motor, episodic memory and mirror neuron regions (i.e., SR processing) compared to controls. This may reflect BPD participants' excessive attempts at understanding self and other, that are however less grounded in sensory reality and episodic memory. Furthermore, greater MSA activation in the task in BPD participants related to worse maintenance of self–other representations over a 3-h period, while greater SR activation in controls related to better maintenance (Beeney et al., 2016). Furthermore, when thinking about resolved and unresolved life events, BPD patients showed hyperactivation in the mentalizing network relative to controls, namely in the ACC, mPFC, and TPJ, and in the dorsolateral PFC, a region involved in autobiographical memory, which may also reflect overcompensation as they inefficiently attempt to reconstruct a coherent narrative of life events (Bozzatello et al., 2019). These findings point toward the imbalance in mentalizing networks that is typical for BPD and shows that the lack of integration between these networks hampers their performance in SOD tasks.

Finally, a new body of research from the so-called "second-person" neuroscience investigates brain activations of two individuals who are in interaction using methods such as hyperscanning functional near-infrared spectroscopy (fNIRS), a technique that is well-suited for "real-life" situations because it is less vulnerable to motion (Gvirts & Perlmutter, 2019). They find significant inter-brain neural synchrony (IBS) between social interaction partners, that is, synchronization of their neural systems (Gvirts & Perlmutter, 2019). Greater IBS has been observed with those interaction partners that are "significant" to us in some way, for instance because we are in a relationship with them (Pan, Cheng, Zhang, Li, & Hu, 2017), or because we need to cooperate with them (Liu et al., 2016), and especially during face-to-face compared to other types of interaction (Jiang et al., 2012). Importantly, IBS may facilitate social alignment and attunement between interaction partners (Shamay-Tsoory, Saporta, Marton-Alper, & Gvirts, 2019). This neural synchrony is specifically found in the TPJ and regions of the PFC, suggesting that these serve as "mutual social attention" systems, shifting the individual's attention toward the social partner to tune into the interaction. Crucially, these same regions are routinely implicated in performance in SOD tasks (Eddy, 2016; Heinisch et al., 2012; Santiesteban et al., 2012; Santiesteban, Banissy, Catmur, & Bird, 2015; Sowden &

Catmur, 2015). It should be noted, however, that fNIRS is limited to the outer cortex, so it cannot be excluded that subcortical regions are also important for IBS.

Oxytocin (OT) has been found to relate to social synchrony (Feldman, Braun, & Champagne, 2019) and to neural IBS (Mu, Guo, & Han, 2016). OT may serve to prioritize tuning into certain social interactions while tuning out others by regulating attention to social cues (Feldman et al., 2019). Because oxytocin is closely linked to the attachment system, secure attachment relationships may help us to determine which interactions are safe and rewarding and should be selectively attended to for emotional co-regulation, closeness, and social learning.

Ostensive cues, such as smiling and shared gaze, were found to enhance infant-adult IBS in studies using fNIRS (Leong et al., 2017; Piazza, Hasenfratz, Hasson, & Lew-Williams, 2018), and the level of parent–infant IBS in the dorsolateral PFC and the frontopolar cortex mediated the association between the parent's and the child's emotion regulation (Reindl, Gerloff, Scharke, & Konrad, 2018). These novel findings shed new light on the long-standing emphasis placed on the importance of parent–infant synchrony, joint attention, and the use of ostensive cues in the development of emotion regulation and attachment (Fonagy et al., 2002, 2007).

Furthermore, in adult learner–instructor pairs, greater IBS in the inferior frontal cortex and the left PFC was found to lead to greater attunement of the learner to the instructor (Pan, Novembre, Song, Li, & Hu, 2018) and enhanced efficiency of learning (Davidesco et al., 2019). These findings suggest that effective social learning crucially requires the instructor and the learner to connect with each other, as reflected in instructor-learner IBS. Individuals with BPD often show a reduced capacity for social learning, expressed in high levels of epistemic mistrust and a reduced capacity for change in psychotherapy, and thus may show reduced IBS (Fonagy, Luyten, Allison, & Campbell, 2017).

Indeed, one study of dyads playing a simple joint attention task while in an fMRI scanner showed decreased neural synchronization in the TPJ in dyads consisting of a BPD patient and a healthy control (HC), compared to HC-remitted BPD dyads or HC-HC dyads; this was associated with a history of childhood trauma (Bilek et al., 2017). Although preliminary, this finding suggests that the mechanism for IBS may be disrupted in BPD, potentially explaining why individuals with BPD seem less able to reap the benefits of attuned social interactions for emotion regulation, social learning, and experiencing closeness to others. Furthermore, decreased IBS in BPD may be the neural reflection of their decreased ability to engage in social situations with self–other ambiguity where they have to co-represent both self and other, which is a necessary condition for practicing SOD (de Bézenac et al., 2018). As the mutual social attention system, regulated by oxytocin, also serves to prioritize more significant interactions over other interactions, a deficiency in this aspect may leave individuals with BPD without a "filter" to decide which social partners should be selectively attended to. Furthermore, the decreased ability of individuals with BPD to "tune into" a social interaction may reflect their decreased ability to learn from others and adopt their perspective and might make them seem "hard to reach." However, this is merely a hypothesis, as more research employing this "second-person" approach to self–other processing in BPD is needed.

Conclusions

Impairments in self–other distinction (SOD) are central in several psychiatric disorders and in borderline personality disorder (BPD) in particular. In this chapter, we have argued that impairments in (embodied) mentalizing is associated with a rigid focus on either "self" or "other," which has detrimental effects on capacities for co-regulation of emotion, social learning, and self-stability. New findings from social neuroscience and experimental psychopathology as reviewed in this chapter provide important new insights into our understanding of SOD impairment in psychopathology, which may improve both prevention and treatment efforts.

References

Adler, J. M., Chin, E. D., Kolisetty, A. P., & Oltmanns, T. F. (2012). The distinguishing characteristics of narrative identity in adults with features of borderline personality disorder: An empirical investigation. *Journal of Personality Disorders, 26*, 498–512. https://doi.org/10.1521/pedi.2012.26.4.498

Agrawal, H. R., Gunderson, J., Holmes, B. M., & Lyons-Ruth, K. (2004). Attachment studies with borderline patients: A review. *Harvard Review of Psychiatry, 12*, 94–104. https://doi.org/10.1080/10673220490447218

American Psychiatric Association. (2013). *DSM-5: Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Press.

Ardizzi, M., Ambrosecchia, M., Buratta, L., Ferri, F., Peciccia, M., Donnari, S., … Gallese, V. (2016). Interoception and positive symptoms in schizophrenia. *Frontiers in Human Neuroscience, 10*, 379. https://doi.org/10.3389/fnhum.2016.00379

Allen, J. G. (2012). Restoring mentalizing in attachment relationships: Treating trauma with plain old therapy. *American Psychiatric Publishing*, Inc.

Backasch, B., Sommer, J., Klohn-Saghatolislam, F., Muller, M. J., Kircher, T. T., & Leube, D. T. (2014). Dysconnectivity of the inferior frontal gyrus: Implications for an impaired self-other distinction in patients with schizophrenia. *Psychiatry Research: Neuroimaging, 223*, 202–209. https://doi.org/10.1016/j.pscychresns.2014.05.007

Ball, J. S., & Links, P. S. (2009). Borderline personality disorder and childhood trauma: Evidence for a causal relationship. *Current Psychiatry Reports, 11*, 63–68. https://doi.org/10.1007/s11920-009-0010-4

Bardi, L., Six, P., & Brass, M. (2017). Repetitive TMS of the temporo-parietal junction disrupts participant's expectations in a spontaneous Theory of Mind task. *Social Cognitive and Affective Neuroscience, 12*, 1775–1782. https://doi.org/10.1093/scan/nsx109

Bargh, J. A. (2011). Unconscious thought theory and its discontents: A critique of the critiques. *Social Cognition, 29*, 629–647. https://doi.org/10.1521/soco.2011.29.6.629

Bargh, J. A. (2014). Our unconscious mind. *Scientific American, 310*, 30–37. https://doi.org/10.1038/scientificamerican0114-30

Barnow, S., Ruge, J., Spitzer, C., & Freyberger, H. J. (2005). Temperament und Charakter bei Personen mit Borderline-Personlichkeitsstorung. [Temperament and character in persons with borderline personality disorder]. *Nervenarzt, 76*, 839–848. https://doi.org/10.1007/s00115-004-1810-8

Bateman, A., O'Connell, J., Lorenzini, N., Gardner, T., & Fonagy, P. (2016). A randomised controlled trial of mentalization-based treatment versus structured clinical management for patients with comorbid borderline personality disorder and antisocial personality disorder. *BMC Psychiatry, 16*, 304. https://doi.org/10.1186/s12888-016-1000-9

Beck, A. T., Freeman, A., & Davis, D. D. (2004). *Cognitive therapy of personality disorders*. New York, NY: Guilford Press.

Beeney, J. E., Hallquist, M. N., Ellison, W. D., & Levy, K. N. (2016). Self-other disturbance in borderline personality disorder: Neural, self-report, and performance-based evidence. *Personality Disorders: Theory, Research, and Treatment, 7*, 28–39. https://doi.org/10.1037/per0000127

Beeney, J. E., Lazarus, S. A., Hallquist, M. N., Stepp, S. D., Wright, A. G. C., Scott, L. N., … Pilkonis, P. A. (2019). Detecting the presence of a personality disorder using interpersonal and self-dysfunction. *Journal of Personality Disorders, 33*, 229–248. https://doi.org/10.1521/pedi_2018_32_345

Bekrater-Bodmann, R., Chung, B. Y., Foell, J., Gescher, D. M., Bohus, M., & Flor, H. (2016). Body plasticity in borderline personality disorder: A link to dissociation. *Comprehensive Psychiatry, 69*, 36–44. https://doi.org/10.1016/j.comppsych.2016.05.002

Belsky, D. W., Caspi, A., Arseneault, L., Bleidorn, W., Fonagy, P., Goodman, M., … Moffitt, T. E. (2012). Etiological features of borderline personality related characteristics in a birth cohort of 12-year-old children. *Development and Psychopathology, 24*, 251–265. https://doi.org/10.1017/S0954579411000812

Bender, D. S., Morey, L. C., & Skodol, A. E. (2011). Toward a model for assessing level of personality functioning in DSM-5, part I: A review of theory and methods. *Journal of Personality Assessment, 93*, 332–346. https://doi.org/10.1080/00223891.2011.583808

Bender, D. S., & Skodol, A. E. (2007). Borderline personality as a self-other representational disturbance. *Journal of Personality Disorders, 21*, 500–517. https://doi.org/10.1521/pedi.2007.21.5.500

Bilek, E., Stossel, G., Schafer, A., Clement, L., Ruf, M., Robnik, L., … Meyer-Lindenberg, A. (2017). State-dependent cross-brain information flow in borderline personality disorder. *JAMA Psychiatry, 74*, 949–957. https://doi.org/10.1001/jamapsychiatry.2017.1682

Bird, G., & Viding, E. (2014). The self to other model of empathy: Providing a new framework for understanding empathy impairments in psychopathy, autism, and alexithymia. *Neuroscience and Biobehavioral Reviews, 47*, 520–532. https://doi.org/10.1016/j.neubiorev.2014.09.021

Biro, S., Alink, L. R., Huffmeijer, R., Bakermans-Kranenburg, M. J., & Van, I. M. H. (2017). Attachment quality is related to the synchrony of mother and infant monitoring patterns. *Attachment and Human Development, 19*, 243–258. https://doi.org/10.1080/14616734.2017.1302487

Blankers, M., Koppers, D., Laurenssen, E. M. P., Peen, J., Smits, M. L., Luyten, P., … Dekker, J. J. M. (2019). Mentalization-based treatment versus specialist treatment as usual for borderline personality disorder: Economic evaluation alongside a randomized controlled trial with 36 months follow-up. *Journal of Personality Disorders*. https://doi.org/10.1521/pedi_2019_33_454

Blatt, S. J., & Luyten, P. (2009). A structural-developmental psychodynamic approach to psychopathology: Two polarities of experience across the life span. *Development and Psychopathology, 21*, 793–814. https://doi.org/10.1017/S0954579409000431

Bornovalova, M. A., Hicks, B. M., Iacono, W. G., & McGue, M. (2009). Stability, change, and heritability of borderline personality disorder traits from adolescence to adulthood: A longitudinal twin study. *Development and Psychopathology, 21*, 1335–1353. https://doi.org/10.1017/S0954579409990186

Botvinick, M., & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature, 391*, 756. https://doi.org/10.1038/35784

Bozzatello, P., Morese, R., Valentini, M. C., Rocca, P., Bosco, F., & Bellino, S. (2019). Autobiographical memories, identity disturbance and brain functioning in patients with borderline personality disorder: An fMRI study. *Heliyon, 5*, e01323. https://doi.org/10.1016/j.heliyon.2019.e01323

Brass, M., Ruby, P., & Spengler, S. (2009). Inhibition of imitative behaviour and social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*, 2359–2367. https://doi.org/10.1098/rstb.2009.0066

Braun, N., Debener, S., Spychala, N., Bongartz, E., Soros, P., Muller, H. H. O., & Philipsen, A. (2018). The senses of agency and ownership: A review. *Frontiers in Psychology, 9*, 535. https://doi.org/10.3389/fpsyg.2018.00535

Bungert, M., Koppe, G., Niedtfeld, I., Vollstadt-Klein, S., Schmahl, C., Lis, S., & Bohus, M. (2015). Pain processing after social exclusion and its relation to rejection sensitivity in borderline personality disorder. *PLoS One, 10*, e0133693. https://doi.org/10.1371/journal.pone.0133693

Carlson, E. A., Egeland, B., & Sroufe, L. A. (2009). A prospective investigation of the development of borderline personality symptoms. *Development and Psychopathology, 21*, 1311–1334. https://doi.org/10.1017/S0954579409990174

Carlson, E. A., Sroufe, L. A., & Egeland, B. (2004). The construction of experience: A longitudinal study of representation and behavior. *Child Development, 75*, 66–83. https://doi.org/10.1111/j.1467-8624.2004.00654.x

Chanen, A. M., & Kaess, M. (2012). Developmental pathways to borderline personality disorder. *Current Psychiatry Reports, 14*, 45–53. https://doi.org/10.1007/s11920-011-0242-y

Colle, L., Gabbatore, I., Riberi, E., Borroz, E., Bosco, F. M., & Keller, R. (2019). Mindreading abilities and borderline personality disorder: A comprehensive assessment using the Theory of Mind Assessment Scale. *Psychiatry Research, 272*, 609–617. https://doi.org/10.1016/j.psychres.2018.12.102

Colonnello, V., Chen, F. S., Panksepp, J., & Heinrichs, M. (2013). Oxytocin sharpens self-other perceptual boundary. *Psychoneuroendocrinology, 38*, 2996–3002. https://doi.org/10.1016/j.psyneuen.2013.08.010

de Aquino Ferreira, L. F., Queiroz Pereira, F. H., Neri Benevides, A. M. L., & Aguiar Melo, M. C. (2018). Borderline personality disorder and sexual abuse: A systematic review. *Psychiatry Research, 262*, 70–77. https://doi.org/10.1016/j.psychres.2018.01.043

de Bézenac, C. E., Swindells, R. A., & Corcoran, R. (2018). The necessity of ambiguity in self-other processing: A psychosocial perspective with implications for mental health. *Frontiers in Psychology, 9*, 2114. https://doi.org/10.3389/fpsyg.2018.02114

Darda, K. M., & Ramsey, R. (2019). The inhibition of automatic imitation: A meta-analysis and synthesis of fMRI studies. *Neuroimage, 197*, 320–329. https://doi.org/10.1016/j.neuroimage.2019.04.059

Davidesco, I., Laurent, E., Valk, H., West, T., Dikker, S., Milne, C., & Poeppel, D. (2019). Brain-to-brain synchrony predicts long-term memory retention more accurately than individual brain measures. *bioRxiv*. https://doi.org/10.1101/644047

De Meulemeester, C., Lowyck, B., Vermote, R., Verhaest, Y., & Luyten, P. (2017). Mentalizing and interpersonal problems in borderline personality disorder: The mediating role of identity diffusion. *Psychiatry Research, 258*, 141–144. https://doi.org/10.1016/j.psychres.2017.09.061

De Panfilis, C., Riva, P., Preti, E., Cabrino, C., & Marchesi, C. (2015). When social inclusion is not enough: Implicit expectations of extreme inclusion in borderline personality disorder. *Personality Disorders: Theory, Research and Treatment, 6*, 301–309. https://doi.org/10.1037/per0000132

Debbané, M., Salaminios, G., Luyten, P., Badoud, D., Armando, M., Solida Tozzi, A., … Brent, B. K. (2016). Attachment, neurobiology, and mentalizing along the psychosis continuum. *Frontiers in Human Neuroscience, 10*, 406. https://doi.org/10.3389/fnhum.2016.00406

Decety, J., & Sommerville, J. A. (2003). Shared representations between self and other: A social cognitive neuroscience view. *Trends in Cognitive Sciences, 7*, 527–533. https://doi.org/10.1016/j.tics.2003.10.004

Deckers, J. W., Lobbestael, J., van Wingen, G. A., Kessels, R. P., Arntz, A., & Egger, J. I. (2015). The influence of stress on social cognition in patients with borderline personality disorder. *Psychoneuroendocrinology, 52*, 119–129. https://doi.org/10.1016/j.psyneuen.2014.11.003

Diamond, L. M., Fagundes, C. P., & Butterworth, M. R. (2012). Attachment style, vagal tone, and empathy during mother-adolescent interactions. *Journal of Research on Adolescence, 22*, 165–184. https://doi.org/10.1111/j.1532-7795.2011.00762.x

Distel, M. A., Rebollo-Mesa, I., Willemsen, G., Derom, C. A., Trull, T. J., Martin, N. G., & Boomsma, D. I. (2009). Familial resemblance of borderline personality disorder features: Genetic or cultural transmission? *PLoS One, 4*, e5334. https://doi.org/10.1371/journal.pone.0005334

Domes, G., Winter, B., Schnell, K., Vohs, K., Fast, K., & Herpertz, S. C. (2006). The influence of emotions on inhibitory functioning in borderline personality disorder. *Psychological Medicine, 36*, 1163–1172. https://doi.org/10.1017/S0033291706007756

Dugue, L., Merriam, E. P., Heeger, D. J., & Carrasco, M. (2018). Specific visual subregions of TPJ mediate reorienting of spatial attention. *Cerebral Cortex, 28*, 2375–2390. https://doi.org/10.1093/cercor/bhx140

Dunn, B. D., Stefanovitch, I., Evans, D., Oliver, C., Hawkins, A., & Dalgleish, T. (2010). Can you feel the beat? Interoceptive awareness is an interactive function of anxiety- and depression-specific symptom dimensions. *Behaviour Research and Therapy, 48*, 1133–1138. https://doi.org/10.1016/j.brat.2010.07.006

Dziobek, I., Preissler, S., Grozdanovic, Z., Heuser, I., Heekeren, H. R., & Roepke, S. (2011). Neuronal correlates of altered empathy and social cognition in borderline personality disorder. *Neuroimage, 57*, 539–548. https://doi.org/10.1016/j.neuroimage.2011.05.005

Eddy, C. M. (2016). The junction between self and other? Temporo-parietal dysfunction in neuropsychiatry. *Neuropsychologia, 89*, 465–477. https://doi.org/10.1016/j.neuropsychologia.2016.07.030

Eggart, M., Lange, A., Binser, M. J., Queri, S., & Muller-Oerlinghausen, B. (2019). Major depressive disorder is associated with impaired interoceptive accuracy: A systematic review. *Brain Sciences, 9*, E131. https://doi.org/10.3390/brainsci9060131

Eshkevari, E., Rieger, E., Longo, M. R., Haggard, P., & Treasure, J. (2012). Increased plasticity of the bodily self in eating disorders. *Psychological Medicine, 42*, 819–828. https://doi.org/10.1017/S0033291711002091

Feldman, R. (2012). Parent-infant synchrony: A biobehavioral model of mutual influences in the formation of affiliative bonds. *Monographs of the Society for Research in Child Development, 77*, 42–51. https://doi.org/10.1111/j.1540-5834.2011.00660.x

Feldman, R., Braun, K., & Champagne, F. A. (2019). The neural mechanisms and consequences of paternal caregiving. *Nature Reviews. Neuroscience, 20*, 205–224. https://doi.org/10.1038/s41583-019-0124-6

Fonagy, P., Gergely, G., Jurist, E. L., & Target, M. (2002). *Affect regulation, mentalization, and the development of the self*. New York, NY: Other Press.

Fonagy, P., Gergely, G., & Target, M. (2007). The parent-infant dyad and the construction of the subjective self. *Journal of Child Psychology and Psychiatry, 48*, 288–328. https://doi.org/10.1111/j.1469-7610.2007.01727.x

Fonagy, P., & Luyten, P. (2016). A multilevel perspective on the development of borderline personality disorder. In D. Cicchetti (Ed.), *Developmental psychopathology. Vol. 3: Maladaptation and psychopathology* (3rd ed., pp. 726–792). New York, NY: Wiley.

Fonagy, P., Luyten, P., & Allison, E. (2015). Epistemic petrification and the restoration of epistemic trust: A new conceptualization of borderline personality disorder and its psychosocial treatment. *Journal of Personality Disorders, 29*, 575–609. https://doi.org/10.1521/pedi.2015.29.5.575

Fonagy, P., Luyten, P., Allison, E., & Campbell, C. (2017). What we have changed our minds about: Part 2. Borderline personality disorder, epistemic trust and the developmental significance of social communication. *Borderline Personality Disorder and Emotion Dysregulation, 4*, 9. https://doi.org/10.1186/s40479-017-0062-8

Fotopoulou, A., & Tsakiris, M. (2017). Mentalizing homeostasis: The social origins of interoceptive inference. *Neuropsychoanalysis, 19*, 3–28. https://doi.org/10.1080/15294145.2017.1294031

Fuchs, T. (2007). Fragmented selves: Temporality and identity in borderline personality disorder. *Psychopathology, 40*, 379–387. https://doi.org/10.1159/000106468

Gallagher, I. I. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences, 4*, 14–21. https://doi.org/10.1016/S1364-6613(99)01417-5

Gergely, G., & Watson, J. S. (1999). Early social-emotional development: Contingency perception and the social biofeedback model. In P. Rochat (Ed.), *Early social cognition* (pp. 101–137). Hillsdale, NJ: Erlbaum.

Giardina, A., Caltagirone, C., & Oliveri, M. (2011). Temporo-parietal junction is involved in attribution of hostile intentionality in social interactions: An rTMS study. *Neuroscience Letters, 495*, 150–154. https://doi.org/10.1016/j.neulet.2011.03.059

Gunderson, J. G., & Lyons-Ruth, K. (2008). BPD's interpersonal hypersensitivity phenotype: A gene-environment-developmental model. *Journal of Personality Disorders, 22*, 22–41. https://doi.org/10.1521/pedi.2008.22.1.22

Gvirts, H. Z., & Perlmutter, R. (2019). What guides us to neurally and behaviorally align with anyone specific? A neurobiological model based on fNIRS hyperscanning studies. *Neuroscientist, 26*(2), 108–116. https://doi.org/10.1177/1073858419861912

Haas, B. W., & Miller, J. D. (2015). Borderline personality traits and brain activity during emotional perspective taking. *Personality Disorders: Theory, Research, and Treatment, 6*, 315–320. https://doi.org/10.1037/per0000130

Han, S., & Northoff, G. (2009). Understanding the self: A cultural neuroscience approach. *Progress in Brain Research, 178*, 203–212. https://doi.org/10.1016/S0079-6123(09)17814-7

Harari, H., Shamay-Tsoory, S. G., Ravid, M., & Levkovitz, Y. (2010). Double dissociation between cognitive and affective empathy in borderline personality disorder. *Psychiatry Research, 175*, 277–279. https://doi.org/10.1016/j.psychres.2009.03.002

Hauschild, S., Winter, D., Thome, J., Liebke, L., Schmahl, C., Bohus, M., & Lis, S. (2018). Behavioural mimicry and loneliness in borderline personality disorder. *Comprehensive Psychiatry, 82*, 30–36. https://doi.org/10.1016/j.comppsych.2018.01.005

Heinisch, C., Kruger, M. C., & Brune, M. (2012). Repetitive transcranial magnetic stimulation over the temporoparietal junction influences distinction of self from famous but not unfamiliar others. *Behavioral Neuroscience, 126*, 792–796. https://doi.org/10.1037/a0030581

Holm, A. L., & Severinsson, E. (2008). The emotional pain and distress of borderline personality disorder: A review of the literature. *International Journal of Mental Health Nursing, 17*, 27–35. https://doi.org/10.1111/j.1447-0349.2007.00508.x

Horowitz, M. J. (2015). Effects of trauma on sense of self. *Journal of Loss and Trauma, 20*, 189–193. https://doi.org/10.1080/15325024.2014.897578

Igelstrom, K. M., Webb, T. W., Kelly, Y. T., & Graziano, M. S. (2016). Topographical organization of attentional, social, and memory processes in the human temporoparietal cortex. *eNeuro, 3*, ENEURO.0060-0016.2016. https://doi.org/10.1523/eneuro.0060-16.2016

Jardri, R., Pins, D., Lafargue, G., Very, E., Ameller, A., Delmaire, C., & Thomas, P. (2011). Increased overlap between the brain areas involved in self-other distinction in schizophrenia. *PLoS One, 6*, e17500. https://doi.org/10.1371/journal.pone.0017500

Jiang, J., Dai, B., Peng, D., Zhu, C., Liu, L., & Lu, C. (2012). Neural synchronization during face-to-face communication. *Journal of Neuroscience, 32*, 16064–16069. https://doi.org/10.1523/jneurosci.2926-12.2012

Jørgensen, C. R. (2006). Disturbed sense of identity in borderline personality disorder. *Journal of Personality Disorders, 20*, 618–644. https://doi.org/10.1521/pedi.2006.20.6.618

Jørgensen, C. R., Berntsen, D., Bech, M., Kjolbye, M., Bennedsen, B. E., & Ramsgaard, S. B. (2012). Identity-related autobiographical memories and cultural life scripts in patients with borderline personality disorder. *Consciousness and Cognition, 21*, 788–798. https://doi.org/10.1016/j.concog.2012.01.010

Kahl, S., & Kopp, S. (2018). A predictive processing model of perception and action for self-other distinction. *Frontiers in Psychology, 9*, 2421. https://doi.org/10.3389/fpsyg.2018.02421

Kanske, P. (2018). The social mind: Disentangling affective and cognitive routes to understanding others. *Interdisciplinary Science Reviews, 43*, 115–124. https://doi.org/10.1080/03080188.2018.1453243

Kanske, P., Schulze, L., Dziobek, I., Scheibner, H., Roepke, S., & Singer, T. (2016). The wandering mind in borderline personality disorder: Instability in self- and other-related thoughts. *Psychiatry Research, 242*, 302–310. https://doi.org/10.1016/j.psychres.2016.05.060

Kaplan, R. A., Enticott, P. G., Hohwy, J., Castle, D. J., & Rossell, S. L. (2014). Is body dysmorphic disorder associated with abnormal bodily self-awareness? A study using the rubber hand illusion. *PLoS One, 9*, e99981. https://doi.org/10.1371/journal.pone.0099981

Kernberg, O. F., & Caligor, E. (2005). A psychoanalytic theory of personality disorders. In M. F. Lenzenweger & J. F. Clarkin (Eds.), *Major theories of personality disorder* (2nd ed., pp. 114–156). New York, NY: Guilford Press.

Klaver, M., & Dijkerman, H. C. (2016). Bodily experience in schizophrenia: Factors underlying a disturbed sense of body ownership. *Frontiers in Human Neuroscience, 10*, 305. https://doi.org/10.3389/fnhum.2016.00305

Krall, S. C., Volz, L. J., Oberwelland, E., Grefkes, C., Fink, G. R., & Konrad, K. (2016). The right temporoparietal junction in attention and social interaction: A transcranial magnetic stimulation study. *Human Brain Mapping, 37*, 796–807. https://doi.org/10.1002/hbm.23068

Lackner, C. L., Bowman, L. C., & Sabbagh, M. A. (2010). Dopaminergic functioning and preschoolers' theory of mind. *Neuropsychologia, 48*, 1767–1774. https://doi.org/10.1016/j.neuropsychologia.2010.02.027

Leong, V., Byrne, E., Clackson, K., Georgieva, S., Lam, S., & Wass, S. (2017). Speaker gaze increases information coupling between infant and adult brains. *Proceedings of the National Academy of Sciences of the United States of America, 114*, 13290–13295. https://doi.org/10.1073/pnas.1702493114

Lieberman, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology, 58*, 259–289. https://doi.org/10.1146/annurev.psych.58.110405.085654

Lind, M., Jorgensen, C. R., Heinskou, T., Simonsen, S., Boye, R., & Thomsen, D. K. (2019). Patients with borderline personality disorder show increased agency in life stories after 12 months of psychotherapy. *Psychotherapy, 56*, 274–284. https://doi.org/10.1037/pst0000184

Liu, N., Mok, C., Witt, E. E., Pradhan, A. H., Chen, J. E., & Reiss, A. L. (2016). NIRS-based hyperscanning reveals inter-brain neural synchronization during cooperative Jenga game with face-to-face communication. *Frontiers in Human Neuroscience, 10*, 82. https://doi.org/10.3389/fnhum.2016.00082

Loffler, A., Foell, J., & Bekrater-Bodmann, R. (2018). Interoception and its interaction with self, other, and emotion processing: Implications for the understanding of psychosocial deficits in borderline personality disorder. *Current Psychiatry Reports, 20*, 28. https://doi.org/10.1007/s11920-018-0890-2

Lombardo, M. V., Chakrabarti, B., & Baron-Cohen, S. (2009). What neuroimaging and perceptions of self-other similarity can tell us about the mechanism underlying mentalizing. *Behavioral and Brain Sciences, 32*, 152–153. https://doi.org/10.1017/S0140525x09000715

Long, M., Verbeke, W., Ein-Dor, T., & Vrticka, P. (2019). A functional neuro-anatomical framework of human attachment: Insights from first and second-person social neuroscience. *Cortex, 126*, 281–321.

Lowyck, B., Luyten, P., Verhaest, Y., Vandeneede, B., & Vermote, R. (2013). Levels of personality functioning and their association with clinical features and interpersonal functioning in patients with personality disorders. *Journal of Personality Disorders, 27*, 320–336. https://doi.org/10.1521/pedi.2013.27.3.320

Luyten, P., & Blatt, S. J. (2013). Interpersonal relatedness and self-definition in normal and disrupted personality development: Retrospect and prospect. *American Psychologist, 68*, 172–183. https://doi.org/10.1037/a0032243

Luyten, P., Campbell, C., & Fonagy, P. (2019). Borderline personality disorder, complex trauma, and problems with self and identity: A social-communicative approach. *Journal of Personality, 88*(1), 88–105. https://doi.org/10.1111/jopy.12483

Luyten, P., & Fonagy, P. (2015). The neurobiology of mentalizing. *Personality Disorders: Theory, Research and Treatment, 6*, 366–379. https://doi.org/10.1037/per0000117

Luyten, P., & Fonagy, P. (2018). The stress-reward-mentalizing model of depression: An integrative developmental cascade approach to child and adolescent depressive disorder based on the Research Domain Criteria (RDoC) approach. *Clinical Psychology Review, 64*, 87–98. https://doi.org/10.1016/j.cpr.2017.09.008

Mayes, L. C. (2006). Arousal regulation, emotional flexibility, medial amygdala function, and the impact of early experience: Comments on the paper of Lewis et al. *Annals of the New York Academy of Sciences, 1094*, 178–192. https://doi.org/10.1196/annals.1376.018

McWilliams, N. (2011). *Psychoanalytic diagnosis: Understanding personality structure in the clinical process* (2nd ed.). New York, NY: Guilford Press.

Meins, E., Fernyhough, C., Fradley, E., & Tuckey, M. (2001). Rethinking maternal sensitivity: Mothers' comments on infants' mental processes predict security of attachment at 12 months. *Journal of Child Psychology and Psychiatry, 42*, 637–648.

Mier, D., Lis, S., Esslinger, C., Sauer, C., Hagenhoff, M., Ulferts, J., … Kirsch, P. (2013). Neuronal correlates of social cognition in borderline personality disorder. *Social Cognitive and Affective Neuroscience, 8*, 531–537. https://doi.org/10.1093/scan/nss028

Mikulincer, M., & Horesh, N. (1999). Adult attachment style and the perception of others: The role of projective mechanisms. *Journal of Personality and Social Psychology, 76*, 1022–1034. https://doi.org/10.1037//0022-3514.76.6.1022

Mikulincer, M., Orbach, I., & Iavnieli, D. (1998). Adult attachment style and affect regulation: Strategic variations in subjective self-other similarity. *Journal of Personality and Social Psychology, 75*, 436–448. https://doi.org/10.1037//0022-3514.75.2.436

Mu, Y., Guo, C., & Han, S. (2016). Oxytocin enhances inter-brain synchrony during social coordination in male adults. *Social Cognitive and Affective Neuroscience, 11*, 1882–1893. https://doi.org/10.1093/scan/nsw106

Muller, L. E., Schulz, A., Andermann, M., Gabel, A., Gescher, D. M., Spohn, A., … Bertsch, K. (2015). Cortical representation of afferent bodily signals in borderline personality disorder: Neural correlates and relationship to emotional dysregulation. *JAMA Psychiatry, 72*, 1077–1086. https://doi.org/10.1001/jamapsychiatry.2015.1252

Neustadter, E. S., Fineberg, S. K., Leavitt, J., Carr, M. M., & Corlett, P. R. (2019). Induced illusory body ownership in borderline personality disorder. *bioRxiv*. https://doi.org/10.1101/628131

New, A. S., aan het Rot, M., Ripoll, L. H., Perez-Rodriguez, M. M., Lazarus, S., Zipursky, E., … Siever, L. J. (2012). Empathy and alexithymia in borderline personality disorder: Clinical and laboratory measures. *Journal of Personality Disorders, 26*, 660–675. https://doi.org/10.1521/pedi.2012.26.5.660

Nolte, T., Bolling, D. Z., Hudac, C. M., Fonagy, P., Mayes, L., & Pelphrey, K. A. (2013). Brain mechanisms underlying the impact of attachment-related stress on social cognition. *Frontiers in Human Neuroscience, 7*, 816. https://doi.org/10.3389/fnhum.2013.00816

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain—A meta-analysis of imaging studies on the self. *Neuroimage, 31*, 440–457. https://doi.org/10.1016/j.neuroimage.2005.12.002

Palmer, C., Ainley, V., & Tsakiris, M. (2019). Fine tuning your heart: A novel method for measuring interoceptive accuracy. *PsyArXiv*. https://doi.org/10.31234/osf.io/qz7r9

Palmer, C. E., & Tsakiris, M. (2018). Going at the heart of social cognition: Is there a role for interoception in self-other distinction? *Current Opinion in Psychology, 24*, 21–26. https://doi.org/10.1016/j.copsyc.2018.04.008

Pan, Y., Cheng, X., Zhang, Z., Li, X., & Hu, Y. (2017). Cooperation in lovers: An fNIRS-based hyperscanning study. *Human Brain Mapping, 38*, 831–841. https://doi.org/10.1002/hbm.23421

Pan, Y., Novembre, G., Song, B., Li, X., & Hu, Y. (2018). Interpersonal synchronization of inferior frontal cortices tracks social interactive learning of a song. *Neuroimage, 183*, 280–290. https://doi.org/10.1016/j.neuroimage.2018.08.005

Pfundmair, M., Rimpel, A., Duffy, K., & Zwarg, C. (2018). Oxytocin blurs the self-other distinction implicitly but not explicitly. *Hormones and Behavior, 98*, 115–120. https://doi.org/10.1016/j.yhbeh.2017.12.016

Piazza, E., Hasenfratz, L., Hasson, U., & Lew-Williams, C. (2018). Infant and adult brains are coupled to the dynamics of natural communication. *bioRxiv*. https://doi.org/10.1101/359810

Prikken, M., van der Weiden, A., Baalbergen, H., Hillegers, M. H., Kahn, R. S., Aarts, H., & van Haren, N. E. (2019). Multisensory integration underlying body-ownership experiences in schizophrenia and offspring of patients: A study using the rubber hand illusion paradigm. *Journal of Psychiatry and Neuroscience, 44*, 177–184. https://doi.org/10.1503/jpn.180049

Reindl, V., Gerloff, C., Scharke, W., & Konrad, K. (2018). Brain-to-brain synchrony in parent-child dyads and the relationship with emotion regulation revealed by fNIRS-based hyperscanning. *Neuroimage, 178*, 493–502. https://doi.org/10.1016/j.neuroimage.2018.05.060

Richetin, J., Preti, E., Costantini, G., & De Panfilis, C. (2017). The centrality of affective instability and identity in borderline personality disorder: Evidence from network analysis. *PLoS One, 12*, e0186695. https://doi.org/10.1371/journal.pone.0186695

Ripoll, L. H., Snyder, R., Steele, H., & Siever, L. J. (2013). The neurobiology of empathy in borderline personality disorder. *Current Psychiatry Reports, 15*, 344. https://doi.org/10.1007/s11920-012-0344-1

Sabbagh, M. A. (2004). Understanding orbitofrontal contributions to theory-of-mind reasoning: Implications for autism. *Brain and Cognition, 55*, 209–219. https://doi.org/10.1016/j.bandc.2003.04.002

Salvatore, J. E., Haydon, K. C., Simpson, J. A., & Collins, W. A. (2013). The distinctive role of romantic relationships in moderating the effects of early caregiving on adult anxious-depressed symptoms over 9 years. *Development and Psychopathology, 25*, 843–856. https://doi.org/10.1017/S0954579413000205

Sansone, R. A., & Sansone, L. A. (2012). Chronic pain syndromes and borderline personality. *Innovations in Clinical Neuroscience, 9*, 10–14.

Santiesteban, I., Banissy, M. J., Catmur, C., & Bird, G. (2012). Enhancing social ability by stimulating right temporoparietal junction. *Current Biology, 22*, 2274–2277. https://doi.org/10.1016/j.cub.2012.10.018

Santiesteban, I., Banissy, M. J., Catmur, C., & Bird, G. (2015). Functional lateralization of temporoparietal junction – imitation inhibition, visual perspective-taking and theory of mind. *European Journal of Neuroscience, 42*, 2527–2533. https://doi.org/10.1111/ejn.13036

Satpute, A. B., & Lieberman, M. D. (2006). Integrating automatic and controlled processes into neurocognitive models of social cognition. *Brain Research, 1079*, 86–97. https://doi.org/10.1016/j.brainres.2006.01.005

Schulze, L., Schmahl, C., & Niedtfeld, I. (2016). Neural correlates of disturbed emotion processing in borderline personality disorder: A multimodal meta-analysis. *Biological Psychiatry, 79*, 97–106. https://doi.org/10.1016/j.biopsych.2015.03.027

Schurz, M., Tholen, M. G., Perner, J., Mars, R. B., & Sallet, J. (2017). Specifying the brain anatomy underlying temporo-parietal junction activations for theory of mind: A review using probabilistic atlases from different imaging modalities. *Human Brain Mapping, 38*, 4788–4805. https://doi.org/10.1002/hbm.23675

Shakoor, S., Jaffee, S. R., Bowes, L., Ouellet-Morin, I., Andreou, P., Happe, F., … Arseneault, L. (2012). A prospective longitudinal study of children's theory of mind and adolescent involvement in bullying. *Journal of Child Psychology and Psychiatry, 53*, 254–261. https://doi.org/10.1111/j.1469-7610.2011.02488.x

Shamay-Tsoory, S. G., & Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: A lesion study. *Neuropsychologia, 45*, 3054–3067. https://doi.org/10.1016/j.neuropsychologia.2007.05.021

Shamay-Tsoory, S. G., Saporta, N., Marton-Alper, I. Z., & Gvirts, H. Z. (2019). Herding brains: A core neural mechanism for social alignment. *Trends in Cognitive Sciences, 23*, 174–186. https://doi.org/10.1016/j.tics.2019.01.002

Silani, G., Lamm, C., Ruff, C. C., & Singer, T. (2013). Right supramarginal gyrus is crucial to overcome emotional egocentricity bias in social judgments. *Journal of Neuroscience, 33*, 15466–15476. https://doi.org/10.1523/JNEUROSCI.1488-13.2013

Smits, M. L., Feenstra, D. J., Eeren, H. V., Bales, D. L., Laurenssen, E. M. P., Blankers, M., … Luyten, P. (2019). Day hospital versus intensive out-patient mentalisation-based treatment for borderline personality disorder: Multicentre randomised clinical trial. *British Journal of Psychiatry, 216*(2), 79–84. https://doi.org/10.1192/bjp.2019.9

Sollberger, D., Gremaud-Heitz, D., Riemenschneider, A., Kuchenhoff, J., Dammann, G., & Walter, M. (2012). Associations between identity diffusion, axis II disorder, and psychopathology in inpatients with borderline personality disorder. *Psychopathology, 45*, 15–21. https://doi.org/10.1159/000325104

Sosic-Vasic, Z., Eberhardt, J., Bosch, J. E., Dommes, L., Labek, K., Buchheim, A., & Viviani, R. (2019). Mirror neuron activations in encoding of psychic pain in borderline personality disorder. *Neuroimage. Clinical, 22*, 101737. https://doi.org/10.1016/j.nicl.2019.101737

Sowden, S., & Catmur, C. (2015). The role of the right temporoparietal junction in the control of imitation. *Cerebral Cortex, 25*, 1107–1113. https://doi.org/10.1093/cercor/bht306

Sowden, S., & Shah, P. (2014). Self-other control: A candidate mechanism for social cognitive function. *Frontiers in Human Neuroscience, 8*, 789. https://doi.org/10.3389/fnhum.2014.00789

Stepp, S. D., Lazarus, S. A., & Byrd, A. L. (2016). A systematic review of risk factors prospectively associated with borderline personality disorder: Taking stock and moving forward. *Personality Disorders: Theory, Research and Treatment, 7*, 316–323. https://doi.org/10.1037/per0000186

Stietz, J., Jauk, E., Krach, S., & Kanske, P. (2019). Dissociating empathy from perspective-taking: Evidence from intra- and inter-individual differences research. *Frontiers in Psychiatry, 10*, 126. https://doi.org/10.3389/fpsyt.2019.00126

Tajadura-Jimenez, A., & Tsakiris, M. (2014). Balancing the "inner" and the "outer" self: Interoceptive sensitivity modulates self-other boundaries. *Journal of Experimental Psychology: General, 143*, 736–744. https://doi.org/10.1037/a0033171

Tomasello, M. (2018). Great apes and human development: A personal history. *Child Development Perspectives, 12*, 189–193. https://doi.org/10.1111/cdep.12281

Tomasello, M., & Vaish, A. (2013). Origins of human cooperation and morality. *Annual Review of Psychology, 64*, 231–255. https://doi.org/10.1146/annurev-psych-113011-143812

Tomova, L., Heinrichs, M., & Lamm, C. (2019). The Other and Me: Effects of oxytocin on self-other distinction. *International Journal of Psychophysiology, 136*, 49–53. https://doi.org/10.1016/j.ijpsycho.2018.03.008

Tsakiris, M. (2017). The multisensory basis of the self: From body to identity to others. *Quarterly Journal of Experimental Psychology, 70*, 597–609. https://doi.org/10.1080/17470218.2016.1181768

Uddin, L. Q., Iacoboni, M., Lange, C., & Keenan, J. P. (2007). The self and social cognition: The role of cortical midline structures and mirror neurons. *Trends in Cognitive Sciences, 11*, 153–157. https://doi.org/10.1016/j.tics.2007.01.001

Uribe, C., Puig-Davi, A., Abos, A., Baggio, H. C., Junque, C., & Segura, B. (2019). Neuroanatomical and functional correlates of cognitive and affective empathy in young healthy adults. *Frontiers in Behavioral Neuroscience, 13*, 85. https://doi.org/10.3389/fnbeh.2019.00085

van der Weiden, A., Prikken, M., & van Haren, N. E. (2015). Self-other integration and distinction in schizophrenia: A theoretical analysis and a review of the evidence. *Neuroscience and Biobehavioral Reviews, 57*, 220–237. https://doi.org/10.1016/j.neubiorev.2015.09.004

van Schie, C. C., Chiu, C. D., Rombouts, S., Heiser, W. J., & Elzinga, B. M. (2019). Stuck in a negative me: fMRI study on the role of disturbed self-views in social feedback processing in borderline personality disorder. *Psychological Medicine, 50*(4), 625–635. https://doi.org/10.1017/S0033291719000448

van Schie, C. C., Chiu, C. D., Rombouts, S. A. R. B., Heiser, W. J., & Elzinga, B. M. (2018). When compliments don't hit but critiques do: An fMRI study into self-esteem and self-knowledge in processing social feedback. *Social Cognitive and Affective Neuroscience, 13*(4), 404–417. https://doi.org/10.1093/scan/nsy014

Villalta, L., Smith, P., Hickin, N., & Stringaris, A. (2018). Emotion regulation difficulties in traumatized youth: A meta-analysis and conceptual review. *European Child and Adolescent Psychiatry, 27*, 527–544. https://doi.org/10.1007/s00787-018-1105-4

Volkert, J., Hauschild, S., & Taubner, S. (2019). Mentalization-based treatment for personality disorders: Efficacy, effectiveness, and new developments. *Current Psychiatry Reports, 21*, 25. https://doi.org/10.1007/s11920-019-1012-5

Wilkinson-Ryan, T., & Westen, D. (2000). Identity disturbance in borderline personality disorder: An empirical investigation. *American Journal of Psychiatry, 157*, 528–541. https://doi.org/10.1176/appi.ajp.157.4.528

Yen, S., Shea, M. T., Sanislow, C. A., Grilo, C. M., Skodol, A. E., Gunderson, J. G., … Morey, L. C. (2004). Borderline personality disorder criteria associated with prospectively observed suicidal behavior. *American Journal of Psychiatry, 161*, 1296–1298. https://doi.org/10.1176/appi.ajp.161.7.1296

Zahavi, D. (2010). Minimal self and narrative self. A distinction in need of refinement. In T. Fuchs, H. Sattel, & P. Heningnsen (Eds.), *The embodied self: Dimensions, coherence, and disorders* (pp. 3–11). Stuttgart, Germany: Schattauer.

Zahavi, D. (2014). *Self and other: Exploring subjectivity, empathy, and shame*. Oxford, UK: Oxford University Press.

Zhao, W., Yao, S., Li, Q., Geng, Y., Ma, X., Luo, L., … Kendrick, K. M. (2016). Oxytocin blurs the self-other distinction during trait judgments and reduces medial prefrontal cortex responses. *Human Brain Mapping, 37*, 2512–2527. https://doi.org/10.1002/hbm.23190

Correction to: Early Theory of Mind Development: Are Infants Inherently Altercentric?

Charlotte Grosse Wiesmann and Victoria Southgate

Correction to:
Chapter 3 in: M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*, https://doi.org/10.1007/978-3-030-51890-5_3

In Chapter 3, an acknowledgement to the funding sources was inadvertently omitted. The correct acknowledgement should be as follows:

Acknowledgement: This work was supported by an ERC Consolidator Grant DEVOMIND 726114 to VS and a Marie Słodowska Curie Fellowship 799734 ToMSelfLink to CGW.

The updated original version of this chapter can be found at
https://doi.org/10.1007/978-3-030-51890-5_3

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_35

Index

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5